Accelerated iterative regularization via dual diagonal descent

Luca Calatroni, Guillaume Garrigos, Lorenzo Rosasco, Silvia Villa

Accelerated iterative regularization via dual diagonal descent

Luca Calatroni[1], Guillaume Garrigos[2], Lorenzo Rosasco[3], and Silvia Villa[4]

[1]Université Côte d'Azur, CNRS, Inria, I3S, France
calatroni@i3s.unice.fr

[2] LPSM, Université de Paris, Sorbonne Université, CNRS, Paris, France
garrigos@lpsm.paris

[3]MaLGa, DIBRIS, Università di Genova, Via Dodecaneso 35, 16146 Genova, Italy
lorenzo.rosasco@unige.it

[4]MaLGa, DIMA, Università di Genova, Via Dodecaneso 35, 16146 Genova, Italy
silvia.villa@unige.it

Abstract

We propose and analyze an accelerated iterative dual diagonal descent algorithm for the solution of linear inverse problems with general regularization and data-fit functions. In particular, we develop an inertial approach of which we analyze both convergence and stability. Using tools from inexact proximal calculus, we prove early stopping results with optimal convergence rates for additive data-fit terms as well as more general cases, such as the Kullback-Leibler divergence, for which different type of proximal point approximations hold.

Keywords. Iterative regularization, duality, acceleration, forward-backward splitting, diagonal methods, stability and convergence analysis.

AMS Classification: 90C25, 49N45, 49N15, 68U10,90C06.

1 Introduction

We are interested in solving the linear inverse problem:

$$\text{find} \quad \bar{x} \in \mathcal{X} \quad \text{s.t.} \quad A\bar{x} = \bar{y}, \tag{1.1}$$

where $A : \mathcal{X} \to \mathcal{Y}$ is a bounded linear operator between two Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$, and $\bar{y} \in \mathcal{Y}$ can be seen as a given measurement of some unknown $\bar{x} \in \mathcal{X}$ we want to recover. In general, the

inverse problem (1.1) is ill-posed as its solution (if it exists) may lack some fundamental properties like uniqueness or stability. A standard approach is assuming the desired $\bar{x}$ to be well-approximated by [38, 25]:

$$\text{find } x^\dagger \in \operatorname{argmin} \left\{ R(x) \text{ s.t. } x \in \operatorname*{argmin}_{x' \in \mathcal{X}} \ell(Ax'; \bar{y}) \right\}, \tag{$\boldsymbol{P_0(\bar{y})}$}$$

called the primal problem in the following. Here, R is a regularization function enforcing some a-priori knowledge on the desired solution $\bar{x}$, while $\ell : \mathcal{Y}^2 \to \mathbb{R} \cup \{+\infty\}$ is a data-fit function. In practice, the data is subject to noise due to, e.g., possible transmission and/or acquisition problems. As a consequence, we only have access to an inexact version $\hat{y}$ of $\bar{y}$. Using $\hat{y}$ in $\boldsymbol{(P_0(\bar{y}))}$ does no longer provide a suitable solution and a standard approach is Tikhonov regularization:

$$\text{find } \hat{x}_\lambda \in \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ p_\lambda(x) := R(x) + \frac{1}{\lambda} \ell(Ax; \hat{y}) \right\}. \tag{$\boldsymbol{P_\lambda(\hat{y})}$}$$

Intuitively, the (so called) regularization parameter $\lambda > 0$ balances the trust in the data $\hat{y}$ with the regularization effect enforced by R. Finding a good approximate solution based on noisy data requires two steps. First, problem ($\boldsymbol{P_\lambda(\hat{y})}$) is solved for various choices of λ by means of a suitable optimization algorithm. Second, the computed solutions are compared using some validation criterion (e.g. discrepancy principles [38], SURE [52, 36], cross-validation [53]) and an optimal parameter λ^* is computed along with the corresponding solution $\hat{x}_{\lambda^*}$. The above procedure is in general very costly computationally and so called iterative regularization methods, that we study in this paper, can provide accurate and more efficient alternatives [38, 20, 41].

Iterative regularization Iterative regularization approaches find an approximation of $x^\dagger$ by running an iterative algorithm and 'stopping' it when close to $x^\dagger$, [18, 29, 30, 26]. The number of iterations plays the role of the regularization parameter, controlling at the same time accuracy and computations. Roughly speaking, an iterative regularization method can be described as follows:

1. Given any data $y \in \mathcal{Y}$, the algorithm generates a sequence $(x_k(y))_{k \in \mathbb{N}}$ using R, ℓ and A.
2. Given the true data $\bar{y}$, the sequence $(x_k(\bar{y}))_{k \in \mathbb{N}}$ solves the hierarchical optimization problem $\boldsymbol{(P_0(\bar{y}))}$, i.e. converges as $k \to \infty$ to the solution $x^\dagger$ of $\boldsymbol{(P_0(\bar{y}))}$.
3. For a given noise level $\delta > 0$ and noisy data $\hat{y}$ such that $\|\bar{y} - \hat{y}\| \leqslant \delta$, there is a stopping time $k(\delta) \in \mathbb{N}$ such that
$$\|\hat{x}_{k(\delta)}(\hat{y}) - x^\dagger\| = O(\delta^\alpha), \text{ for some } \alpha > 0. \tag{1.2}$$

The quantity $O(\delta^\alpha)$ is often called the *rate* of the considered regularization method, and the exponent α quantifies its efficiency.

Previous results For quadratic data-fit terms ℓ and square-norm regularization terms R, both Tikhonov and iterative regularization (such as the Landweber algorithm) have been shown to be optimal, in the sense that their reconstruction error (1.2) has optimal rate $O(\delta^{\frac{1}{2}})$ [38]. Optimal results with possibly fewer iterations are also known to be possible considering accelerated approaches [38, 48]. For quadratic data-fit, and general strongly convex regularizers an iterative

regularization procedure combined with a Morozov-type discrepancy principle was also shown to be optimal in [30], and accelerated approaches based on a dual accelerated gradient descent was shown to be optimal with less iterations in [45]. Iterative regularization methods have been studied also for convex regularizers in [29] where estimates in terms of Bregman distance were proved (see, e.g., [30, 18] for Tikhonov-type approaches), but no explicit rates in the form (1.2) were shown. More general iterative algorithms defined in Banach spaces have been studied in [42, 43, 28] for linear and non-linear inverse problems and in [26] for L^1 and Total Variation (TV) regularization. For results on iterative regularization for data-fit terms other than squared norm, we mention [24] for results in the framework of Bregman distances and [39] where a dual diagonal descent **(3D)** algorithm is considered. To the best of our knowledge, accelerated iterative regularization approaches have not been studied in this general setting.

Contribution and organization of the paper In this paper, we study a novel accelerated iterative regularization algorithm for strongly convex regularization terms and general data-fits and prove its optimality. As for the quadratic case [48, 45], we show that acceleration can lead to optimal results in much fewer iterations. Our approach, dubbed **(I3D)** , extends the **(3D)** iterative algorithm studied in [39] introducing an inertial term yielding acceleration. From an optimization perspective, the rationale behind these results is that inertial dynamics are able to exploit information in previous iterates to converge faster to an optimal solution. However, as pointed out in [37], differently from basic schemes, inertial methods suffer from errors accumulation that need to be controlled along the iterations and balanced with the improvements observed in the convergence speed.
The paper is organized as follows. In Section 2 we introduce the main notations and assumptions. In Section 3 we introduce and analyze a diagonal inertial dynamic in continuous time. In Section 4, we derive **(I3D)** as a discretization of the continuous dynamic, and we study its convergence and stability properties. In Section 6 the results are illustrated for a number of specific data-fit terms including Kullback-Leibler divergence data-fit.

2 Main assumptions and background on diagonal methods

We begin fixing the notation. Let $\mathcal{H}$ be a Hilbert space with scalar product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$. Given $y \in \mathcal{H}$, and $\varrho \in \mathbb{R}_+$ and $\mathbb{B}(y, \varrho)$ is the open ball of center y and radius ϱ. we denote by $\Gamma_0(\mathcal{H})$ the set of proper, convex and lower semi-continuous functions from $\mathcal{H}$ to $]-\infty, +\infty]$. We say that $f \in \Gamma_0(\mathcal{H})$ is σ-strongly convex if $f - \sigma\|\cdot\|^2/2 \in \Gamma_0(\mathcal{H})$, with $\sigma \in]0, +\infty[$. We recall that the subdifferential of $f \in \Gamma_0(\mathcal{H})$ is the multi-valued operator $\partial f : \mathcal{H} \to 2^{\mathcal{H}}$ defined by setting

$$(\forall x \in \mathcal{H}) \quad \partial f(x) := \left\{ u \in \mathcal{H} : f(x') - f(x) - \langle u, x' - x \rangle \geqslant 0 \text{ for all } x' \in \mathcal{H} \right\}. \tag{2.1}$$

Note that if f is Gateaux differentiable at $x \in \mathcal{H}$, then $\partial f(x) = \{\nabla f(x)\}$, see, e.g. , [21, Proposition 17.31 i)]. We also recall that for all $x \in \mathcal{H}$ and $\tau > 0$ the proximity operator $\mathrm{prox}_{\tau f} : \mathcal{H} \to \mathcal{H}$ of $f \in \Gamma_0(\mathcal{H})$ with parameter τ is defined by:

$$\mathrm{prox}_{\tau f}(x) = (I + \partial f)^{-1}(x) = \operatorname*{argmin}_{x' \in \mathcal{H}} \left\{ f(x') + \frac{1}{2\tau}\|x' - x\|^2 \right\}.$$

Given $f \in \Gamma_0(\mathcal{H})$, we will denote by $f^* : \mathcal{H} \to [-\infty, +\infty]$ the Fenchel conjugate of f

$$(\forall u \in \mathcal{H}) \quad f^*(u) := \sup_{x \in \mathcal{H}} \{\langle u, x \rangle_{\mathcal{H}} - f(x)\}.$$

The Fenchel conjugate of f belongs to $\Gamma_0(\mathcal{H})$, and if f is σ-strongly convex then f^* is differentiable at any point, with a σ^{-1}-Lipschitz continuous gradient, see, e.g. [21, Theorem 18.15]. Furthermore, the following property holds, see [21, Theorem 16.23]:

$$(\forall (x, u) \in \mathcal{H}^2) \quad u \in \partial f(x) \Leftrightarrow x \in \partial f^*(u).$$

Finally, given two real sequences $(a_k)_{k \geqslant 1}$ and $(b_k)_{k \geqslant 1}$, we will write $a_k = O(b_k)$ whenever there exists a positive constant $M > 0$ such that $a_k \leqslant M b_k$ for all $k \geqslant 1$. We will use the more precise notation $a_k = \Theta(b_k)$ if both conditions $a_k = O(b_k)$ and $b_k = O(a_k)$ hold. Note also that we will use the same notation $\| \cdot \|$ and $\langle \cdot, \cdot \rangle$ for the norm and the scalar product in every Hilbert space we consider.

2.1 Main assumptions

We make the following assumptions on the data-fit ℓ and the regularizer R:

- (L_1) For all $y \in \mathcal{Y}$, $\ell_y := \ell(\cdot, y) \in \Gamma_0(\mathcal{Y})$ and it is coercive.
- (L_2) For all $(y_1, y_2) \in \mathcal{Y}^2$, $\ell(y_1, y_2) = 0 \iff y_1 = y_2$.
- (L_3) there exists $q \in [1, +\infty[$, $\varrho \in]0, +\infty]$, $\gamma \in]0, +\infty[$ such that

$$\forall y \in \mathbb{B}(\bar{y}, \varrho), \quad \frac{\gamma}{q} \|y - \bar{y}\|^q \leqslant \ell(y, \bar{y}).$$

- (R_1) R is σ-strongly convex, with $\sigma \in]0, +\infty[$,
- (R_2) $\partial R(x^\dagger) \cap \operatorname{Im} A^* \neq \emptyset$.

If assumption (L_3) is satisfied, we say that $\ell_{\bar{y}}$ is locally q-conditioned at $\bar{y}$. If (L_3) is satisfied with $\rho = +\infty$, then we say that $\ell_{\bar{y}}$ is q-conditioned at $\bar{y}$. These assumptions cover a wide range of inverse problems, as discussed next.

Definition 2.1 A data-fit is additive if there exists $\mathcal{N} \in \Gamma_0(\mathcal{Y})$ such that

$$(\forall (y_1, y_2) \in \mathcal{Y}^2) \quad \ell(y_1, y_2) = \mathcal{N}(y_1 - y_2).$$

Example 2.2 (Data-fit functions) For $\mathcal{Y} = \mathbb{R}^d$, the additive data-fit functions defined by the functions $\mathcal{N}$ below trivially satisfy (L_1)-(L_2). In addition, $\ell_{\bar{y}}$ satisfies (L_3) if and only if $\mathcal{N}$ is locally conditioned at 0, and this is the case, as we show below:

- $\mathcal{N}(y) = \frac{1}{2}\|y\|^2$ is 2-conditioned at 0, with $\gamma = 1$.

- $\mathcal{N}(y) = \frac{1}{q}\|y\|_q^q$, for $q \geqslant 1$, is q-conditioned at 0 with $\gamma = d^r$, where $r := \min(\frac{1}{q} - \frac{1}{2}, 0)$. Note that this includes the case of the ℓ^1-norm.
- the weighted sum [40] $\mathcal{N}(y) = \alpha\|y\|_1 + \frac{1}{2}\|y\|_2^2$, for $\alpha > 0$, is 1-conditioned at 0, with $\gamma = \alpha$.
- the Huber data-fit function [32] $\mathcal{N}(y) = \sum_{i=1}^{d} h_\nu(y^i)$, where $h_\nu \colon \mathbb{R} \to \mathbb{R}_+$ is the Huber smoothing function, defined for $\nu > 0$ as

$$(\forall t \in \mathbb{R}) \quad h_\nu(t) := \begin{cases} \frac{1}{2\nu}t^2 & \text{if } |t| \leqslant \nu \\ |t| - \frac{\nu}{2} & \text{otherwise.} \end{cases}$$

  The Huber data-fit is locally 2-conditioned at 0. It is enough to choose $\varrho \in]0, +\infty[$ with $\gamma = \min\{1/\nu, 2\varrho - \nu/\varrho^2\}$.
- the exact penalization $\mathcal{N}(y) = 0$ if $y = 0$, $\mathcal{N}(y) = +\infty$ otherwise, is 1-conditioned with $\gamma = 1$.

We also mention a non-additive data-fit function used in several applications:

- the Kullback-Leibler divergence, defined as:

$$\ell(y_2, y_1) = \mathrm{KL}(y_1, y_2) := \sum_{i=1}^{d} \mathrm{kl}(y_1^i, y_2^i), \tag{2.2}$$

  where

$$(\forall (t_1, t_2) \in \mathbb{R}^2) \quad \mathrm{kl}(t_1, t_2) := \begin{cases} t_1 \log \dfrac{t_1}{t_2} - t_1 + t_2 & \text{if } (t_1, t_2) \in]0, +\infty[^2, \\ +\infty & \text{otherwise.} \end{cases}$$

  The Kullback-Leibler divergence is locally 2-conditioned at $\bar{y}$ for every $\varrho \in]0, +\infty[$, with $\gamma = \frac{2}{\varrho c^2} + \frac{2}{\varrho^2 c} \ln \frac{c}{\varrho + c}$, with $c = d\|\bar{y}\|_\infty$ (see Lemma A.2).

Example 2.3 (Regularizers) A regularizer widely used in signal/image processing as a sparsifying prior is the ℓ^1-norm of the coefficients with respect to an orthonormal basis, or a more general dictionary. Another popular choice in imaging is the total variation semi-norm [49], due to its ability to preserve edges, together with its generalizations [27, 34]. For some specific tasks in computer vision and machine learning, there is a need for structured sparsity. This kind of prior can be enforced with the use of group sparsity inducing norms [56, 17]. While not being strongly convex, these regularizers can be included in our framework by adding a strongly convex term $\frac{\sigma}{2}\|\cdot\|^2$ where σ is small positive parameter, in the flavor of the elastic net regularization [58].

2.2 Iterative methods based on continuous and discrete dynamics

It is useful to review some approaches to solve (1.1), the hierarchical problem $(\boldsymbol{P_0(\bar{y})})$ and the Tikhonov-regularized problem $(\boldsymbol{P_\lambda(\hat{y})})$. In particular, we focus on approaches based on duality and/or combined with diagonal dynamics.

Mirror descent approaches A class of methods to solve (1.1) consider the problem

$$\text{find } x^\dagger \in \operatorname*{argmin}_{x\in\mathcal{X}} \{R(x) + \delta_{\bar{y}}(Ax)\},$$

where the constraint (1.1) is encoded by the indicator function $\delta_{\bar{y}}$. Using Fenchel-Rockafeller duality, we derive the corresponding dual problem:

$$\text{find } u^\dagger \quad \text{s.t.} \quad u^\dagger \in \operatorname*{argmin}_{u\in\mathcal{Y}} \{d_0(u) := R^*(-A^*u) + \langle \bar{y}, u\rangle\}. \tag{D_0}$$

Since R^* is smooth (see (ii) in Lemma A.1), a gradient method can be used to solve **(D_0)**, see [45]. This coincides, up to a change of variables, with mirror descent approaches [22] and linearized Bregman iterations [30, 18], where R plays the role of the mirror function. How to extend this approach to solve **($P_0(\bar{y})$)** is not clear.

Primal diagonal dynamics A classical approach to solve hierarchical problems like **($P_0(\bar{y})$)** is the *diagonal principle*, based on the fact that when $\hat{y} = \bar{y}$ and $\lambda \to 0$, problem **($P_\lambda(\hat{y})$)** converges towards **($P_0(\bar{y})$)** in an appropriate sense [4, Theorem 2.6]. In this view, diagonal approaches have been considered as non-autonomous dynamics solving **($P_\lambda(\hat{y})$)** with a parameter λ monotonically decreasing to zero. The simplest example of a continuous diagonal dynamic is the diagonal steepest descent differential inclusion with initial $t_0 > 0$ defined by

$$x(t_0) = x_0, \quad \lambda(t) \searrow 0, \quad \dot{x}(t) + \partial p_{\lambda(t)}(x(t)) \ni 0, \tag{PD_λ}$$

where $p_{\lambda(t)}(x(t))$ is defined in **($P_\lambda(\hat{y})$)**. This dynamic is studied in [10, 12, 6] where convergence of $x(t)$ to $x^\dagger$ was guaranteed provided that $\lambda(t) \to 0$ *fast enough*, namely $\lambda \in L^{1/(q-1)}([t_0, +\infty))$, where $q \in [1, +\infty)$ is the exponent in (L_3), see [6, Corollary 3.3, Remark 4.4]. Discrete counterparts of **(PD_λ)** have also been studied [19, 11, 35]. They can be seen as a variant of the Forward-Backward algorithm applied to solve problem **($P_\lambda(\hat{y})$)**, where the penalization parameter tends to zero along the iterations. A main drawback of this type of algorithms is that they are expansive for non-smooth data-fit terms, since they require to compute the proximal operator of $\ell_{\bar{y}} \circ A$. A possible way to overcome this issue is applying Fenchel-Rockafellar duality to **($P_\lambda(\hat{y})$)**. The main advantage of considering the dual problem **(D_λ)** is that the linear operator appears there only in composition with the smooth function R^*. Then it is possible to apply an explicit gradient step to $R^* \circ (-A^*)$, while the non-smooth data-fit term can be cheaply treated via its proximal operator.

Dual diagonal dynamics The dual problem of **($P_\lambda(\hat{y})$)** is

$$\text{find } u_\lambda \in \operatorname*{argmin}_{u\in\mathcal{Y}} \left\{ d_\lambda(u) := R^*(-A^*u) + \frac{1}{\lambda}\ell^*(\lambda u; \hat{y}) \right\}. \tag{D_λ}$$

Solutions of **(D_λ)** are related to those of **($P_\lambda(\hat{y})$)** via the formula $x_\lambda = \nabla R^*(-A^*u_\lambda)$, which holds thanks to the strong convexity of R. A natural question is whether the diagonal principle can be applied on to the dual problem **(D_λ)** as well. The corresponding dual diagonal continuous dynamics read

$$u(t_0) = u_0, \quad \lambda(t) \searrow 0, \quad \begin{cases} x(t) = \nabla R^*(-A^*u(t)), \\ \dot{u}(t) + \partial d_{\lambda(t)}(u(t)) \ni 0. \end{cases}, \tag{DD_λ}$$

where, similarly as before, provided that $\lambda \in L^{1/(q-1)}([t_0, +\infty))$, the trajectory $x(t)$ is guaranteed to converge to $x^\dagger$. The discrete counterpart of **(DD_λ)** is called Dual Diagonal Descent algorithm **(3D)** , and its convergence and stability properties are studied in [39]. For $\hat{y} \in \mathcal{Y}$ such that $\|\hat{y} - \bar{y}\| \leqslant \delta$ and additive data-fit functions, the authors showed that stopping the algorithm at $k_\delta = \Theta(k^{-2/3})$ guarantees that (1.2) holds with $\alpha = 1/3$. However, that this rate is not optimal [38]. As we show in the following, in this paper our approach recover the optimal rate, while providing the benefits of accelerated approaches, namely an earlier stopping time.

3 Continuous inertial dual diagonal dynamic

First-order inertial algorithms are popular in optimization due to their faster convergence on smooth and non-smooth convex problems, see e.g. [47, 23]. In several papers continuous inertial dynamics have been studied considering appropriate Lyapunov functions [54, 44, 3]. As already noted, their regularization properties are also known for quadratic data-fit terms [48, 45].

Next, we propose an inertial approach for general data-fit terms, introducing a variant of the dynamic in **(DD_λ)**. For a given $\alpha > 0$ and initial $t_0 > 0$, let

$$(u(t_0), \dot{u}(t_0)) = (u_0, \dot{u}_0), \quad \lambda(t) \searrow 0, \quad \begin{cases} x(t) = \nabla R^*(-A^* u(t)), \\ \ddot{u}(t) + \dfrac{\alpha}{t}\dot{u}(t) + \partial d_{\lambda(t)}(u(t)) \ni 0. \end{cases} \tag{IDD_λ}$$

The asymptotic behavior of the trajectories of this inertial differential inclusion will be analyzed next, while its discrete counterpart will be studied in the rest of the paper. We first add one remark.

Remark 3.1 The idea of coupling inertia with Tikhonov regularization is not new. In [8] an inertial variant of the primal dynamic **(PD_λ)** is proposed for $R = \|\cdot\|^2/2$. The corresponding inertial primal diagonal approach is:

$$(x(t_0), \dot{x}(t_0)) = (x_0, \dot{x}_0), \quad \lambda(t) \searrow 0, \quad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \lambda(t)\partial p_{\lambda(t)}(x(t)) \ni 0. \tag{IPD_λ}$$

Under a suitable decay assumption on $\lambda(\cdot)$ the authors guarantee fast convergence and regularization [8, Section 6]. Compared to **(IPD_λ)**, in our dual formulation **(IDD_λ)** we take advantage of a different scaling between the data-fit and the regularizer. Indeed, in **(IDD_λ)** the data-fit is multiplied by $\lambda(t)^{-1} \to +\infty$, while in **($IPD_\lambda$)** the regularizer is multiplied by $\lambda(t) \to 0$. For first-order systems this difference is essentially cosmetic, the two approaches being equivalent, for an appropriate change of variables [12]. However, for second-order systems these two scalings describe different dynamics [13, Section 4]. This difference can be understood looking at the limits (in the Γ-convergence sense) of the corresponding parametrized functions, indeed,

$$\text{if } \lambda \searrow 0, \quad p_\lambda \to p_0 := R + \delta_{\operatorname{argmin} \ell_y \circ A} \quad \text{and} \quad \lambda p_\lambda \to \delta_{\operatorname{dom} R} + \ell_y \circ A.$$

3.1 Convergence of the continuous inertial dual diagonal dynamic

Next, we study the convergence properties of the trajectories of **(IDD_λ)**, assuming their existence to simplify the analysis. We remark that if d_λ is assumed to be gradient-Lipschitz continuous, global

existence and uniqueness results of a classical $C^2([t_0,+\infty),\mathbb{R}_+)$ solution to $(\boldsymbol{IDD_\lambda})$ hold from the Cauchy-Lipschitz theorem. However, this assumption requires the data-fidelity function $\ell_{\bar{y}}$ to be strongly convex (see [21, Theorem 18.15]), which is in general not the case for most of the data-fit terms, see Example 2.2. We refer to [31, 3] for further details. In the following Theorem, we show that the inertial term $(\boldsymbol{IDD_\lambda})$ ensures that the dual function values $d_{\lambda(t)}(u(t))$ tend to $\inf d_0$ at a $O(t^{-2})$ rate as expected for inertial methods. Further, switching from the dual to the primal problem by means of $x(t)=\nabla R^*(-A^*u(t))$, we prove the convergence of $x(\cdot)$ to $x^\dagger$. To prove these results, assumptions on the decay of $\lambda(\cdot)$ are needed, as it is usual for dynamics such as those in $(\boldsymbol{PD_\lambda})$ and $(\boldsymbol{IPD_\lambda})$. For instance, in [12, 6], it is shown that the trajectories of $(\boldsymbol{PD_\lambda})$ converge under the condition $\lambda \in L^{\frac{1}{q-1}}([t_0,+\infty))$ [6]. Similarly, we consider the following assumption:

(Λ) $\lambda : [0,+\infty[\to]0,+\infty[$ is a non-increasing differentiable function such that $\lim_{t\to\infty}\lambda(t)=0$. Moreover, if q defined in assumption (L_3) is strictly greater than 1, the quantity $\Lambda_c = \int_{t_0}^{+\infty} t\lambda^{\frac{1}{q-1}}(t)\,dt$ is finite.

Remark 3.2 A sufficient condition ensuring the validity of (Λ) is that $\lambda(\cdot)\in L^{\frac{1}{2(q-1)}}([t_0,+\infty))$, see Lemma A.4 in the Appendix.

We are now ready to state the main convergence result for continuous dynamics. Note that Lemma A.1(iii) ensures that the set of solutions of problem $(\boldsymbol{D_0})$ is nonempty. To prove fast convergence results of the dual function values, we follow the approach considered in [7, 54, 3] and define a suitable Lyapunov-type function.

Theorem 3.3 *Let the assumptions (L_1)-(L_3), (R_1)-(R_2), (Λ) hold true. Let $u^\dagger \in \operatorname{argmin} d_0$ and assume that $\lambda(t_0)\|u^\dagger\| \leqslant \gamma\varrho^{q-1}/q$. Let $\alpha \geqslant 3$ and let the pair $(x(\cdot),u(\cdot))$ be a solution to $(\boldsymbol{IDD_\lambda})$ in the following sense:*

- *$u \in \mathcal{C}^1([t_0,+\infty[,\mathcal{Y})$, and $x=\nabla R^*\circ(-A^*)\circ u$,*
- *for every $T>t_0$, $\dot{u}$ and $d_{\lambda(\cdot)}\circ u$ are absolutely continuous on $[t_0,T]$,*
- *for a.e. $t\in[t_0,+\infty[$, $-\ddot{u}(t)-\frac{\alpha}{t}\dot{u}(t)\in\partial d_{\lambda(t)}(u(t))$.*

Then, there exists an explicit $C\in]0,+\infty[$ such that

$$\forall t>t_0 \quad d_{\lambda(t)}(u(t))-\inf d_0 \leqslant \frac{C}{t^2} \quad \textit{and} \quad \|x(t)-x^\dagger\| \leqslant \frac{\sqrt{2C}}{\sqrt{\sigma}t}.$$

Proof. Define the following energy:

$$(\forall t\geqslant t_0)\quad \mathcal{E}(t) := t^2\left(d_{\lambda(t)}(u(t))-\inf d_0\right)+\frac{1}{2}\|(\alpha-1)(u(t)-u^\dagger)+t\dot{u}(t)\|^2.$$

From now on we will use the following shorthand notation:

$$R_A^* := R^*\circ(-A^*), \qquad \ell_{\bar{y}}^*(\cdot) := \ell(\cdot,\bar{y})^* \tag{3.1}$$

so that the composite dual function d_λ can be written as $d_\lambda(u) = R_A^*(u) + \lambda^{-1}\ell_{\bar{y}}^*(\lambda u)$, for every $u \in \mathcal{Y}$. Since $\partial d_{\lambda(t)}(u(t)) = \nabla R_A^*(u(t)) + \partial \ell_{\bar{y}}^*(\lambda(t)u(t))$ [21, Proposition 16.6 and Corollary 16.53], the notion of solution introduced entails that there exists some $\eta : [t_0, +\infty) \to \mathcal{Y}$ such that

$$\text{for a.e. } t > t_0, \quad \ddot{u}(t) + \frac{\alpha}{t}\dot{u}(t) + \nabla R_A^*(u(t)) + \eta(t) = 0 \ \text{ and } \ \eta(t) \in \partial \ell_{\bar{y}}^*(\lambda(t)u(t)).$$

We divide the proof in two steps.

Step 1. Fast convergence rates The function $\mathcal{E}$ is differentiable a.e. on $[t_0, +\infty[$ since it is absolutely continuous. We thus compute its derivative and obtain:

$$\begin{aligned}\dot{\mathcal{E}}(t) = 2t\Big(d_{\lambda(t)}(u(t)) - \inf d_0\Big) + \frac{t^2\dot{\lambda}(t)}{\lambda^2(t)}\Big(\langle \eta(t), \lambda(t)u(t)\rangle - \ell_{\bar{y}}^*(\lambda(t)u(t))\Big) \\ + t^2\langle \dot{u}(t), \ddot{u}(t) + \frac{\alpha}{t}\dot{u}(t) + \nabla R_A^*(u(t)) + \eta(t)\rangle + t(\alpha-1)\langle u(t) - u^\dagger, \frac{\alpha}{t}\dot{u}(t) + \ddot{u}(t)\rangle.\end{aligned}$$

The second term in the expression above is non-positive because λ is differentiable and decreasing and, moreover, by convexity of $\ell_{\bar{y}}(\cdot)$ together with Lemma A.1(ii), there holds

$$\ell_{\bar{y}}^*(\lambda(t)u(t)) - \langle \eta(t), \lambda(t)u(t)\rangle \leqslant \ell_{\bar{y}}^*(0) = 0.$$

Furthermore, the third term is equal to zero a.e. since $u(\cdot)$ is a solution of $(\boldsymbol{IDD_\lambda})$ by assumption. We thus deduce that for a.e. $t > t_0$

$$\dot{\mathcal{E}}(t) \leqslant 2t\Big(d_{\lambda(t)}(u(t)) - \inf d_0\Big) + t(\alpha-1)\langle u^\dagger - u(t), -\ddot{u}(t) - \frac{\alpha}{t}\dot{u}(t)\rangle. \tag{3.2}$$

Using that $-\ddot{u}(t) - \frac{\alpha}{t}\dot{u}(t) \in \partial d_{\lambda(t)}(u(t))$ and from the convexity of $d_{\lambda(t)}(\cdot)$ we have:

$$\text{for a.e. } t > t_0 \quad \langle u^\dagger - u(t), -\ddot{u}(t) - \frac{\alpha}{t}\dot{u}(t)\rangle \leqslant d_{\lambda(t)}(u^\dagger) - d_{\lambda(t)}(u(t)).$$

We now add and subtract $\inf d_0 = d_0(u^\dagger)$ and define $r_{\lambda(t)}(u^\dagger) := d_{\lambda(t)}(u^\dagger) - \inf d_0$. We get:

$$\langle u^\dagger - u(t), -\ddot{u}(t) - \frac{\alpha}{t}\dot{u}(t)\rangle \leqslant r_{\lambda(t)}(u^\dagger) + \inf d_0 - d_{\lambda(t)}(u(t)).$$

Applying this inequality to (3.2), since $\alpha \geqslant 3$ and $d_{\lambda(t)}(u(t)) - \inf d_0 \geqslant 0$ (see Proposition A.1.(iv)), we get:

$$\dot{\mathcal{E}}(t) \leqslant t(3-\alpha)\Big(d_{\lambda(t)}(u(t)) - \inf d_0\Big) + t(\alpha-1)r_{\lambda(t)}(u^\dagger) \leqslant t(\alpha-1)r_{\lambda(t)}(u^\dagger). \tag{3.3}$$

To bound the right hand side, we now apply Lemma A.1(vi) and deduce that, since $\lambda(t) \leqslant \lambda(t_0)$:

$$\dot{\mathcal{E}}(t) \leqslant c(\alpha-1)t\lambda(t)^{\frac{1}{q-1}},$$

where the constant c is defined as:

$$c := \begin{cases} 0 & \text{if } q = 1, \\ (1-(1/q))\gamma^{-1/(q-1)}\|u^\dagger\|^{q/(q-1)} & \text{if } q > 1, \end{cases} \tag{3.4}$$

and it is finite in both cases. Since the above inequality holds for a.e. $t > t_0$, assumption (Λ) yields that for a.e. $t > t_0$:

$$\mathcal{E}(t) = \mathcal{E}(t_0) + \int_{t_0}^t \dot{\mathcal{E}}(t) \leqslant \mathcal{E}(t_0) + c(\alpha-1)\Lambda_c.$$

Defining $C := \mathcal{E}(t_0) + c(\alpha-1)\Lambda_c$, we derive

$$d_{\lambda(t)}(u(t)) - \inf d_0 \leqslant \frac{C}{t^2}. \tag{3.5}$$

Step 2. Convergence rate for the primal iterates From (3.5) in combination to Lemma A.1(v), we get

$$\begin{aligned}\frac{\sigma}{2}\|x(t)-x^\dagger\|^2 &\leqslant d_0(u(t)) - \inf d_0 = (d_0(u(t)) - d_{\lambda(t)}(u(t))) + (d_{\lambda(t)}(u(t)) - \inf d_0)\\ &\leqslant (d_0(u(t)) - d_{\lambda(t)}(u(t))) + \frac{C}{t^2}.\end{aligned}$$

The monotonicity property of Lemma A.1(iv) implies that the first term on the right hand side above is non-positive, whence we get

$$\|x(t)-x^\dagger\| \leqslant \frac{\sqrt{2C}}{\sqrt{\sigma}t}.$$

□

4 Inertial Dual Diagonal Descent (I3D) Algorithm

In this section, we study the convergence properties of the discrete analogue of **(IDD_λ)**, deriving an accelerated version of the **(3D)** algorithm in [39].

4.1 From the continuous dynamic to the discrete algorithm

We follow a standard approach of computiong the time-discretization of the continuous dynamical system [1, 15, 54, 7]. Recalling the notation in (3.1), we note that **(IDD_λ)** can be equivalently written as

$$\begin{cases} x(t) = \nabla R^*(-A^*u(t)),\\ \ddot{u}(t) + \frac{\alpha}{t}\dot{u}(t) + \partial\ell^*_{\bar{y}}(\lambda(t)u(t)) + \nabla R^*_A(u(t)) \ni 0. \end{cases} \tag{4.1}$$

We discretize (4.1) explicitly with respect to the smooth component ∇R^*_A and semi-implicitly with respect to the non-smooth one $\partial\ell^*_{\bar{y}}$. In other words, we discretize implicitly the trajectories, while leaving explicit the dependence on the discretized values λ_k. For $k \geqslant 0$, a fixed time step-size $h > 0$ and for some time discretization points $t_k = kh$, we set $u_k := u(t_k)$, $\lambda_k := \lambda(t_k)$ and derive the finite difference scheme:

$$\begin{cases} x_k = \nabla R^*(-A^*u_k),\\ \frac{1}{h^2}(u_{k+1} - 2u_k + u_{k-1}) + \frac{\alpha}{kh^2}(u_k - u_{k-1}) + \partial\ell^*_{\bar{y}}(\lambda_k u_{k+1}) + \nabla R^*_A(w_k) \ni 0, \end{cases}$$

where w_k is a linear combination of u_k and u_{k-1}, made clear in the following. After straightforward calculations, we rewrite the system above as

$$\begin{cases} x_k = \nabla R^*(-A^*u_k),\\ u_{k+1} + h^2\partial\ell^*_{\bar{y}}(\lambda_k u_{k+1}) \ni u_k + \left(1 - \frac{\alpha}{k}\right)(u_k - u_{k-1}) - h^2\nabla R^*_A(w_k). \end{cases}$$

Hence, by setting $\alpha_k = 1 - \alpha/k$, $\tau := h^2$ and $w_k := u_k + \alpha_k(u_k - u_{k-1})$, we get

$$\begin{cases} w_k &= u_k + \alpha_k(u_k - u_{k-1}),\\ u_{k+1} &= \left(I + \frac{\tau}{\lambda_k}\partial\ell^*_{\bar{y}}(\lambda_k\cdot)\right)^{-1}(w_k - \tau\nabla R^*_A(w_k)),\\ x_{k+1} &= \nabla R^*(-A^*u_{k+1}). \end{cases}$$

Note that the proximal operator of the map $\ell_{\bar{y}}^*(\lambda_k \cdot)$ with parameter τ/λ_k appears, in combination with an explicit gradient step for R_A^*. We can thus introduce the Inertial Dual Diagonal Descent **(I3D)** algorithm

$$u_0 = u_1 \in \mathcal{Y}, \quad \text{compute for } k \geqslant 1 \quad \begin{cases} w_k &= u_k + \alpha_k(u_k - u_{k-1}), \\ u_{k+1} &= \operatorname{prox}_{\frac{\tau}{\lambda_k}\ell_{\bar{y}}^*(\lambda_k \cdot)}\left(w_k - \tau \nabla R_A^*(w_k)\right), \\ x_{k+1} &= \nabla R^*(-A^* u_{k+1}). \end{cases} \qquad \textbf{(I3D)}$$

This algorithm depends on three parameters: the stepsize $\tau > 0$, the relaxation parameters $(\lambda_k)_k$ and the friction parameters $(\alpha_k)_k$. The stepsize will be chosen depending on the value of the Lipschitz constant of ∇R_A^*. For the relaxation parameters, we will consider a discrete analogue of the assumption (Λ) formulated in the continuous setting. For the friction parameters α_k, we will allow more general values not necessarily corresponding to $\alpha_k = 1 - \alpha/k$ as above. We gather the requirements on the parameters in the following assumptions:

- (P_1) $\tau \in \left(0, \frac{\sigma^2}{\|A\|^2}\right]$, where $\sigma > 0$ is defined in assumption (R_1).
- (P_2) α_k is non-negative and for every $k \geqslant 1$, $t_k := 1 + \sum_{i=k}^{+\infty} \prod_{j=k}^{i} \alpha_j$ is finite, with $t_k = \Theta(k)$.
- (P_3) (λ_k) is a strictly positive non-increasing sequence such that $\lim_{k\to\infty} \lambda_k = 0$. Moreover, by defining
$$\Lambda := \begin{cases} \sum_{k\geqslant 1} t_{k+1} \lambda_k^{1/(q-1)} & \text{if } q > 1, \\ 0 & \text{if } q = 1, \end{cases} \tag{4.2}$$
we have that $\Lambda < +\infty$.
- (P_4) For some $u^\dagger \in \operatorname{argmin} d_0$, we have $\lambda_0 \|u^\dagger\| \leqslant \gamma \varrho^{q-1}/q$.

Remark 4.1 (On assumption (P_3)) As commented in Remark 3.2, one can check that a sufficient condition for (P_3) to hold is that $\lambda \in \ell^{\frac{1}{2(q-1)}}(\mathbb{N})$. In particular, if we consider a sequence verifying $\lambda_k = O\left(k^{-\theta}\right)$ for some $\theta > 0$, it is easy to verify that (P_3) holds as long as $\theta > 2(q-1)$. For $q = 1$, (for instance if $\ell(y_1, y_2) = \|y_1 - y_2\|_1$), no summability condition is required. Roughly speaking, the assumption $\lambda \in \ell^{\frac{1}{2(q-1)}}(\mathbb{N})$ means that $\lambda \in \ell^\infty(\mathbb{N})$, which is already implied by $\lim_{k\to\infty} \lambda_k = 0$.

Remark 4.2 (On assumption (P_4)) For many choices of data-fits, $\varrho = +\infty$ (see Example 2.2), in which case the assumption is automatically satisfied. Also, note that in assumption (P_3), we require λ_k to tend to zero. This means that $\lambda_K \|u^\dagger\| \leqslant \gamma \varrho^{q-1}/q$ for some $K \in \mathbb{N}$. In this case, up to a time rescaling $k \leftarrow k + K$ our estimates hold true.

Following [5], we require the sequence of friction parameters (α_k) to satisfy (P_2), a particular summability property guaranteeing a technical condition crucial in the following. We summarize such a requirement and the resulting condition in the following lemma.

Lemma 4.3 ([5, Lemma 2.1]) *Assume that (α_k) is non-negative and satisfies*

$$\sum_{i=k}^{+\infty} \prod_{j=k}^{i} \alpha_j < +\infty, \quad \textit{for every } k \geqslant 1. \tag{4.3}$$

Then, the sequence defined by

$$t_k := 1 + \sum_{i=k}^{+\infty} \prod_{j=k}^{i} \alpha_j \tag{4.4}$$

is well-defined, and satisfies for every $k \geqslant 1$ the following properties:

$$1 + \alpha_k t_{k+1} = t_k, \qquad t_{k+1}^2 - t_k^2 \leqslant t_{k+1}. \tag{4.5}$$

Remark 4.4 (Classical choices of α_k and t_k) Definitions (4.3) and (4.4) above accommodate standard choices of sequences (α_k) and (t_k). For example, in his seminal work Nesterov [46] considered

$$\alpha_k = \frac{t_k - 1}{t_{k+1}} \qquad \text{and} \qquad t_{k+1} = \frac{\sqrt{1 + 4t_k^2} + 1}{2}, \qquad t_1 = 1, \tag{4.6}$$

which can be shown to verify the two conditions (4.3) and (4.4), as well as $k/2 \leqslant t_k \leqslant k$. For a given $\alpha > 1$, the two asymptotically equivalent choices

$$\alpha_k = 1 - \frac{\alpha}{k}, \quad t_{k+1} = \frac{k}{\alpha - 1}, \quad \text{and} \quad \alpha_k = \frac{k-1}{k+\alpha-1}, \quad t_{k+1} = \frac{k+\alpha-1}{\alpha-1}$$

have been recently considered in [33, 2, 9] and can be shown to satisfy (P_2). For $\alpha = 3$, these sequences are asymptotically equivalent to Nesterov sequences (4.6).

Remark 4.5 (Splitting of the loss) In [39] the decomposition $\ell_{\bar{y}} = \phi_{\bar{y}} \,\square\, \psi_{\bar{y}}$ was considered, where $\square$ is the infimal convolution and $\psi_{\bar{y}}$ is the possible strongly convex component of $\ell_{\bar{y}}$. In this case, the dual function $\ell_{\bar{y}}^*(\cdot)$ can be expressed as $\ell_{\bar{y}}^* = \psi_{\bar{y}}^* + \phi_{\bar{y}}^*$, where $\phi_{\bar{y}}^*$ is in general non-smooth, while $\phi_{\bar{y}}^*$ has Lipschitz gradient and can therefore be incorporated with the smooth term R_A^* in the dual function d_λ. For several data discrepancies, however, $\psi_{\bar{y}} = \delta_{\{0\}}$ (see [39, Section 4.3]). To simplify the presentation, we do not consider this decomposition.

4.2 Fast convergence of the algorithm

We now prove the discrete analogue of Theorem 3.3 for ***(I3D)***. We follow the approach considered in [14, 54, 7, 5].

Theorem 4.6 (Fast convergence) *Let the assumptions (L_1)-(L_3), (R_1)-(R_2), (P_1)-(P_4) hold true. Let (x_k) and (u_k) be the sequences generated by algorithm **(I3D)**. Then, there exists $C \in \,]0, +\infty[$ such that*

$$d_{\lambda_k}(u_k) - \inf d_0 \leqslant \frac{C}{t_k^2} \quad \text{and} \quad \|x_k - x^\dagger\| \leqslant \frac{\sqrt{2C}}{\sqrt{\sigma} t_k}. \tag{4.7}$$

Proof. Let $u^\dagger \in \operatorname{argmin} d_0$ be the minimizer of d_0 defined in assumption (P_4), and define, for every $k \geqslant 1$, the discrete Lyapunov energy function:

$$\mathcal{E}(k) := t_k^2 \Big(d_{\lambda_k}(u_k) - \inf d_0 \Big) + \frac{1}{2\tau} \|z_k - u^\dagger\|^2, \tag{4.8}$$

where z_k is defined as:

$$z_k := u_{k-1} + t_k(u_k - u_{k-1}). \tag{4.9}$$

Our goal is to get an estimate on the decay of $\mathcal{E}$ along time. More precisely, we will show that for every $k \geqslant 1$

$$\mathcal{E}(k+1) - \mathcal{E}(k) \leqslant t_{k+1}\Big(d_{\lambda_k}(u^\dagger) - \inf d_0\Big), \tag{4.10}$$

which can be seen as a discrete analogue of (3.3), and from which the desired accelerated convergence rates will follow in a straightforward manner.

For simplicity, let us denote by ℓ_k the function defined by setting

$$(\forall u \in \mathcal{Y}) \quad \ell_k(u) = \lambda_k^{-1}\ell_{\bar{y}}^*(\lambda_k u).$$

To prove (4.10), we define for every $k \geqslant 1$ the operator $G_k : \mathcal{Y} \to \mathcal{Y}$ as

$$G_k(z) := \frac{1}{\tau}\Big(z - \operatorname{prox}_{\tau\ell_k}(z - \tau\nabla R_A^*(z))\Big)$$

and notice that the proximal step of ***(I3D)*** can be written in terms of G_k as $u_{k+1} = w_k - \tau G_k(w_k)$. The descent lemma yields (see, e.g., [5, 33])

$$d_{\lambda_k}(w - \tau G_k(w)) \leqslant d_{\lambda_k}(u) + \langle G_k(w), w - u\rangle - \frac{\tau}{2}\|G_k(w)\|^2, \quad \text{for all } w, u \in \mathcal{Y}, \tag{4.11}$$

Evaluating (4.11) for $u = u_k$ and $w = w_k$, we get

$$d_{\lambda_k}(u_{k+1}) \leqslant d_{\lambda_k}(u_k) + \langle G_k(w_k), w_k - u_k\rangle - \frac{\tau}{2}\|G_k(w_k)\|^2. \tag{4.12}$$

Evaluating (4.11) for $u = u^\dagger$ and $w = w_k$ we derive

$$d_{\lambda_k}(u_{k+1}) \leqslant d_{\lambda_k}(u^\dagger) + \langle G_k(w_k), w_k - u^\dagger\rangle - \frac{\tau}{2}\|G_k(w_k)\|^2. \tag{4.13}$$

We now multiply (4.12) by $t_{k+1} - 1$ and we add it to (4.13), and we obtain

$$\begin{aligned} t_{k+1}d_{\lambda_k}(u_{k+1}) \leqslant{}& (t_{k+1}-1)d_{\lambda_k}(u_k) + d_{\lambda_k}(u^\dagger) \\ &+ \langle G_k(w_k), (t_{k+1}-1)(w_k - u_k) + (w_k - u^\dagger)\rangle - \frac{\tau}{2}t_{k+1}\|G_k(w_k)\|^2. \end{aligned} \tag{4.14}$$

As an immediate consequence of Lemma 4.3, we observe the following property:

$$\begin{aligned} (t_{k+1}-1)(w_k - u_k) + w_k &= u_k + t_{k+1}(w_k - u_k) \\ &= u_k + t_{k+1}\alpha_k(u_k - u_{k-1}) \\ &= u_{k-1} + (1 + t_{k+1}\alpha_k)(u_k - u_{k-1}) \\ &= u_{k-1} + t_k(u_k - u_{k-1}) = z_k. \end{aligned}$$

Thanks to (4.9), the fact that $z_k - \tau t_{k+1}G_k(w_k) = z_{k+1}$ and the previous equality, we can reorder the terms in (4.14) and rewrite it as

$$\begin{aligned} t_{k+1}(d_{\lambda_k}(u_{k+1}) - d_{\lambda_k}(u^\dagger)) \leqslant{}& (t_{k+1}-1)(d_{\lambda_k}(u_k) - d_{\lambda_k}(u^\dagger)) \\ &+ \frac{1}{2\tau t_{k+1}}\left(\|z_k - u^\dagger\|^2 - \|z_{k+1} - u^\dagger\|^2\right). \end{aligned}$$

We now multiply everything by t_{k+1}, re-arrange and get

$$\begin{aligned} &t_{k+1}^2(d_{\lambda_k}(u_{k+1}) - d_{\lambda_k}(u^\dagger)) + \frac{1}{2\tau}\|z_{k+1} - u^\dagger\|^2 \\ &\leqslant (t_{k+1}^2 - t_{k+1})(d_{\lambda_k}(u_k) - d_{\lambda_k}(u^\dagger)) + \frac{1}{2\tau}\|z_k - u^\dagger\|^2. \end{aligned}$$

Equivalently:

$$\begin{aligned} &t_{k+1}^2\Big(d_{\lambda_k}(u_{k+1}) - d_{\lambda_k}(u^\dagger)\Big) + \frac{1}{2\tau}\|z_{k+1} - u^\dagger\|^2 \\ &\leqslant t_k^2\Big(d_{\lambda_k}(u_k) - d_{\lambda_k}(u^\dagger)\Big) + (t_{k+1}^2 - t_{k+1} - t_k^2)\Big(d_{\lambda_k}(u_k) - d_{\lambda_k}(u^\dagger)\Big) + \frac{1}{2\tau}\|z_k - u^\dagger\|^2. \end{aligned}$$

To get the desired terms, we first use on the left-hand side the monotonicity property of the function $d_{\lambda_k}(\cdot)$ as a function of k (see Lemma A.1(iv)) and then add and subtract in the parentheses the term $\inf d_0$, thus getting:

$$\begin{aligned} &t_{k+1}^2\Big(d_{\lambda_{k+1}}(u_{k+1}) - \inf d_0\Big) + \frac{1}{2\tau}\|z_{k+1} - u^\dagger\|^2 \\ &\leqslant t_k^2\Big(d_{\lambda_k}(u_k) - \inf d_0\Big) + (t_{k+1}^2 - t_{k+1} - t_k^2)\Big(d_{\lambda_k}(u_k) - \inf d_0\Big) \\ &+ t_{k+1}\Big(d_{\lambda_k}(u^\dagger) - \inf d_0\Big) + \frac{1}{2\tau}\|z_k - u^\dagger\|^2. \end{aligned}$$

After rearranging and using the definition of $\mathcal{E}$ in (4.8), we deduce:

$$\mathcal{E}(k+1) + (t_k^2 + t_{k+1} - t_{k+1}^2)\Big(d_{\lambda_k}(u_k) - \inf d_0\Big) \leqslant \mathcal{E}(k) + t_{k+1}\Big(d_{\lambda_k}(u^\dagger) - \inf d_0\Big).$$

Thanks to (4.5) and Lemma A.1(iv), we can neglect the second term in the left-hand side of the above inequality, finally getting the desired (4.10). Iterating (4.10) recursively gives

$$\mathcal{E}(k) \leqslant \mathcal{E}(1) + \sum_{j=1}^{k-1} t_{j+1}\Big(d_{\lambda_j}(u^\dagger) - \inf d_0\Big). \tag{4.15}$$

To bound the sum appearing on the right hand side, we need to analyze the residuals $r_j := d_{\lambda_j}(u^\dagger) - \inf d_0$. Similarly as for the estimation obtained in the continuous case, we can use for this purpose the property in Lemma A.1(vi) and get that for some fixed constant c independent on j (defined analogously as in (3.4)), we have

$$r_j \leqslant c\lambda_j^{\frac{1}{q-1}}, \qquad \text{for every } j \geqslant 1.$$

By assumption (P_3), with Λ as in (4.2), we thus conclude that

$$\sum_{j=1}^{k-1} t_{j+1} r_j \leqslant c\sum_{j=1}^{k-1} t_{j+1}\lambda_j^{\frac{1}{q-1}} \leqslant c\Lambda < +\infty$$

This allows us to deduce from (4.15) the convergence rate on the dual values in (4.7), by taking $C = \mathcal{E}(1) + c\Lambda$. Finally, the convergence rate on the primal iterates in (4.7) follows from Lemma A.1(v). □

Remark 4.7 (Nesterov scheme as a special case) Let f be any differentiable function in $\Gamma_0(\mathcal{X})$ with a Lipschitz gradient. Take $R = f^*$, $A = -I$, $\bar{y} = 0$, and $\ell(y_1, y_2) = \delta_0(y_2 - y_1)$, so that assumptions (L_1)-(L_3) and (R_1)-(R_2) are verified. In that case, $d_0 = f$, and ***(I3D)*** reads

$$u_0 = u_1 \in \mathcal{Y}, \quad \text{compute for } k \geqslant 1 \quad \begin{cases} w_k &= u_k + \alpha_k(u_k - u_{k-1}), \\ u_{k+1} &= w_k - \tau \nabla f(w_k), \\ x_{k+1} &= \nabla f(u_{k+1}), \end{cases}$$

which in the dual exactly performs Nesterov's method [47]. From our rates and Lemma A.1(iv), we deduce that $f(u_k) - \inf f = O(k^{-2})$. And, according to Nemirovski and Yudin optimality result [47, Theorem 2.1.7], these rates are optimal over the class of Lipschitz smooth convex functions.

Remark 4.8 (Different growth for t_k) In assumption (P_2) we ask the sequence (t_k) to satisfy $t_k = \Theta(k)$, but this is actually not used in the proof of Theorem 4.6. What is crucial is that $t_k < +\infty$, so that Lemma 4.3 can be used. Indeed, one might ask whether it is possible to require $t_k = \Theta(k^\beta)$, with $\beta > 1$ to improve the rates in (3.5). It is a simple exercise to verify that this is not possible, since (4.5) implies $t_k \leqslant t_1 k$, hence $\beta \leqslant 1$ so that the best rates are achieved for $\beta = 1$.

5 Stability properties in the presence of errors

We now study the iterative regularization properties of ***(I3D)*** in the presence of noisy data, given by

$$\hat{u}_0 = \hat{u}_1 \in \mathcal{Y}, \quad \text{compute for } k \geqslant 1 \quad \begin{cases} \hat{w}_k &= \hat{u}_k + \alpha_k(\hat{u}_k - \hat{u}_{k-1}), \\ \hat{u}_{k+1} &= \operatorname{prox}_{\frac{\tau}{\lambda_k}\ell^*_{\hat{y}}(\lambda_k \cdot)}\left(\hat{w}_k - \tau \nabla R^*_A(\hat{w}_k)\right), \\ \hat{x}_{k+1} &= \nabla R^*(-A^* \hat{u}_{k+1}). \end{cases}$$

A first question is how much the dual and primal iterates $\hat{u}_k$ and $\hat{x}_k$ are affected by noise in terms of both convergence and stability. We discuss these issues showing that the noise can be interpreted as an error in the calculation of the proximal step of the algorithm. Before starting, we motivate our analysis with the following example.

Example 5.1 Assume $\mathcal{Y} = \mathbb{R}$ and $\hat{y} = \bar{y} + \delta$, for some $\bar{y},\ \delta > 0$. The algorithm makes use of the datum only in the evaluation of the proximal operator $\operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\hat{y}}(\lambda \cdot)}$. One way to measure the impact of the noise is to find an upper bound for $|\operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\hat{y}}(\lambda \cdot)}(w) - \operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\bar{y}}(\lambda \cdot)}(w)|$, for $w \in \mathcal{Y}$ (see [39, Lemma 10]. As two particular cases consider:

- $\ell_y = \frac{1}{2}|\cdot - y|^2$. We have:

$$\sup_{\bar{y} \in \mathcal{Y}} \sup_{w \in \mathcal{Y}} |\operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\hat{y}}(\lambda \cdot)}(w) - \operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\bar{y}}(\lambda \cdot)}(w)| = \frac{\tau\delta}{1 + \tau\lambda}.$$

- $\ell_y = \mathrm{kl}(y; \cdot)$. We have:

$$\sup_{\bar{y} \in \mathcal{Y}} \sup_{w \in \mathcal{Y}} |\operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\hat{y}}(\lambda \cdot)}(w) - \operatorname{prox}_{\frac{\tau}{\lambda}\ell^*_{\bar{y}}(\lambda \cdot)}(w)| = \sqrt{\frac{\tau\delta}{\lambda}}.$$

In the former example, the error assumed in the evaluation of $\bar{y}$ is propagated along the iterations with order δ. However, a different behavior is observed for the latter example. The square-root dependence on δ makes the estimate worse in a small noise regime, when $\delta \ll 1$. Further, the diagonal convergence to zero of the sequence (λ_k), assumed in (P_3), makes the overall error growing fast along the iterations.

Example 5.1 shows that not all losses behave the same. Our analysis needs to be flexible enough to take these differences into account, and avoid sub-optimal results via a worst-case analysis. This is the purpose of this section, where we will see that additive losses (in the sense of Definition 2.1) behave essentially like $\frac{1}{2}|\cdot - y|^2$, while Kullback-Leibler belongs to a class of less stable losses.

5.1 ε-subdifferentials and inexact proximal calculus

In this section, we define perturbations in a precise way. We first recall standard definitions regarding the approximate subdifferential and proximal-type minimization problems.

Definition 5.2 (ε-subdifferential [57]) Let $\mathcal{H}$ be a Hilbert space, $f \in \Gamma_0(\mathcal{H})$ and $\varepsilon \geqslant 0$. The ε-subdifferential of f at $x \in \operatorname{dom} f$ is the set

$$\partial_\varepsilon f(x) = \left\{ u \in \mathcal{H} : f(x') \geqslant f(x) + \langle u, x' - x \rangle - \varepsilon, \text{ for all } x' \in \mathcal{H} \right\}.$$

Such a notion generalizes that of the subdifferential recalled in (2.1). In particular, if $\varepsilon \geqslant 0$, then $\partial f(x) \subset \partial_\varepsilon f(x)$ for any $x \in \mathcal{H}$, and we have

$$0 \in \partial_\varepsilon f(x) \quad \Longleftrightarrow \quad x \in \operatorname{argmin}^\varepsilon f = \{x' \in \mathcal{H} : f(x') \leqslant \inf f + \varepsilon\}.$$

The following are useful characterizations of the proximal operator of $f \in \Gamma_0(\mathcal{H})$ with parameter $\eta > 0$,

$$p = \operatorname{prox}_{\eta f}(x) \Leftrightarrow \frac{x-p}{\eta} \in \partial f(p) \Leftrightarrow p = \operatorname*{argmin}_z \left\{ f(z) + \frac{1}{2\eta}\|z - x\|^2 \right\}. \tag{5.1}$$

Next, we introduce notions of approximation of proximal points that can be seen as relaxed conditions of the characterizations in (5.1) (for details see [50, 16]).

Definition 5.3 (Approximation of proximal points) Let $f \in \Gamma_0(\mathcal{H})$, $x \in \mathcal{H}$, $\eta > 0$ and $p := \operatorname{prox}_{\eta f}(x)$. We say that $\hat{p} \in \mathcal{H}$ is:

- a **type** 1 approximation of p with precision ε_1, and we write $\hat{p} \approx_1^{\varepsilon_1} p$, if:

$$\exists e \in \mathcal{H},\ \exists (\varepsilon_1, \varepsilon_2, \varepsilon_3) \in [0, +\infty[^2,\ \|e\| \leqslant \varepsilon_3,\ \varepsilon_2^2 + \varepsilon_3^2 \leqslant \varepsilon_1^2,\ \frac{x + e - \hat{p}}{\eta} \in \partial_{\frac{\varepsilon_2^2}{2\eta}} f(\hat{p}).$$

- a **type** 2 approximation of p with precision ε_2, and we write $\hat{p} \approx_2^{\varepsilon_2} p$, if

$$\exists \varepsilon_2 \in [0, +\infty[,\ \frac{x - \hat{p}}{\eta} \in \partial_{\frac{\varepsilon_2^2}{2\eta}} f(\hat{p}).$$

- a **type** 3 approximation of p with precision ε_3, and we write $\hat{p} \approx_3^{\varepsilon_3} p$, if

$$\exists e \in \mathcal{H}, \exists \varepsilon_3 \in [0, +\infty],\ \|e\| \leqslant \varepsilon_3,\ \ \frac{x + e - \hat{p}}{\eta} \in \partial f(\hat{p}).$$

Type 3 approximation simply describes an additive error in the argument of the proximal map, i.e. $\hat{p} = \text{prox}_{\eta f}(x + e)$. We show in Section 6.1 that this type of error arises naturally when additive-type data-fit functions are used. Type 2 approximation considers an error in the subdifferential operator. The definition of type 1 approximation can be seen as a combination of type 2 and 3 approximations, and the following Lemma provides an easy characterization.

Lemma 5.4 ([51, 50]) *Let $f \in \Gamma_0(\mathcal{H})$, $x \in \mathcal{H}$, $\eta > 0$. Then:*

$$\hat{p} \approx_1^{\varepsilon_1} \text{prox}_{\eta f}(x) \quad \Leftrightarrow \quad \hat{p} \in \text{argmin}_{\varepsilon_1} \left\{ f(\cdot) + \frac{1}{2\eta} \| \cdot - x \|^2 \right\}.$$

In summary, those three type of errors correspond to relaxations of the characterizations in (5.1). We are ready to study the stability properties of the **(I3D)** algorithm.

5.2 Stability estimates in the presence of errors

Using the notions introduced in the previous section, we can quantify the error due to replacing $\bar{y}$ by $\hat{y}$. In particular, recalling Definition 5.3, we assume that at each iteration the proximal step with $\hat{y}$ is an i-type approximation of the proximal step with $\bar{y}$, where $i \in \{1, 2, 3\}$.

(E_i) For every $k \geqslant 1$, $\exists \varepsilon_{i,k} \in [0, +\infty[$, s.t. $\forall w \in \mathcal{Y}$,

$$\text{prox}_{\frac{\tau}{\lambda_k} \ell_{\hat{y}}^*(\lambda \cdot)}(w) \approx_i^{\varepsilon_{i,k}} \text{prox}_{\frac{\tau}{\lambda_k} \ell_{\bar{y}}^*(\lambda \cdot)}(w).$$

In Section 6 we show that this is natural for classical data-fit terms. We are ready to prove our second main result for ***(I3D)*** about error estimates under assumption (E_i) with $i = 1$. Stability results for type $2, 3$ approximations are deduced noting that for these choices the error terms with $\varepsilon_{3,k}$ and $\varepsilon_{2,k}$ vanish, respectively, for every k.

Theorem 5.5 (Error estimates for type 1 errors) *Assume that (L_1)-(L_3), (R_1)-(R_2), (P_1)-(P_4) hold true. Let $(\hat{x}_k)$, $(\hat{u}_k)$ be the sequences generated by **(I3D)** with noisy datum $\hat{y}$, and that (E_i) holds with $i = 1$. Then, we have the stability estimate:*

$$(\forall k \geqslant 1) \quad t_k^2 \frac{\sigma\tau}{2} \|\hat{x}_k - x^\dagger\|^2 \leqslant C + \sum_{j=1}^{k-1} t_{j+1}^2 \varepsilon_{2,j}^2 + \frac{5}{2} \Big(\sum_{j=1}^{k-1} t_{j+1} \varepsilon_{3,j} \Big)^2, \tag{5.2}$$

where the constant C is defined as

$$C := \begin{cases} 2\tau t_1^2 \Big(d_1(\hat{u}_0) - \inf d_0 \Big) + \|\hat{u}_0 - u^\dagger\|^2 & \text{if } q = 1, \\ 2\tau t_1^2 \Big(d_1(\hat{u}_0) - \inf d_0 \Big) + \|\hat{u}_0 - u^\dagger\|^2 + 2\tau\Lambda(1 - \frac{1}{q}) \gamma^{-1/(q-1)} \|u^\dagger\|^{q/(q-1)} & \text{if } q > 1. \end{cases}$$

Proof. Following the proof of Theorem 4.6, we define the discrete energy function

$$\hat{\mathcal{E}}(k) := t_k^2 \Big(d_{\lambda_k}(\hat{u}_k) - \inf d_0 \Big) + \frac{1}{2\tau} \|\hat{z}_k - u^\dagger\|^2, \tag{5.3}$$

for $k \geqslant 1$, where $u^\dagger \in \operatorname{argmin} d_0$ (so that $\inf d_0 = d_0(u^\dagger)$) and $\hat{z}_k$ is defined as:

$$\hat{z}_k := \hat{u}_{k-1} + t_k(\hat{u}_k - \hat{u}_{k-1}).$$

Since $\hat{u}_{k+1} \approx_1^{\varepsilon_{1,k}} \operatorname{prox}_{\tau\lambda_k^{-1}\ell_{\hat{y}}^*(\lambda_k\,\cdot)}(\hat{w}_k - \tau\nabla\mathcal{R}_A(\hat{w}_k))$, using Definition 5.3, we have

$$\xi_k := \frac{\hat{w}_k + e_k - \hat{u}_{k+1}}{\tau}, \qquad \xi_k - \nabla\mathcal{R}_A(\hat{w}_k) \in \partial_{\frac{\varepsilon_{2,k}^2}{2\tau}} \ell_{\hat{y}}^*(\lambda_k\hat{u}_{k+1}),$$

where $e_k \in \mathcal{H}$, $\varepsilon_{2,k}^2 + \varepsilon_{3,k}^2 \leqslant \varepsilon_{1,k}^2$ and $\|e_k\| \leqslant \varepsilon_{3,k}$. Without loss of generality, we can assume that $\varepsilon_{2,k}^2 + \varepsilon_{3,k}^2 = \varepsilon_{1,k}^2$. Thus, by the descent lemma in [55, Lemma 4.1] applied to $d_{\lambda_k} = \mathcal{R}_A + \lambda_k^{-1}\ell_{\hat{y}}^*(\lambda_k\,\cdot)$, we derive

$$d_{\lambda_k}(\hat{u}_{k+1}) \leqslant d_{\lambda_k}(u) + \langle \hat{u}_{k+1} - u, \xi_k\rangle + \frac{L}{2}\|\hat{u}_{k+1} - \hat{w}_k\|^2 + \frac{\varepsilon_{2,k}^2}{2\tau}, \qquad \forall u \in \mathcal{Y},$$

where $L = \|A\|^2/\sigma^2$. Using the fact that $\tau L \leqslant 1$, rearranging and neglecting non-positive quantities, we obtain that for all $u \in \mathcal{Y}$:

$$\begin{aligned}
d_{\lambda_k}(\hat{u}_{k+1}) &\leqslant d_{\lambda_k}(u) - \frac{1}{\tau}\|\hat{u}_{k+1} - \hat{w}_k\|^2 + \langle \hat{u}_{k+1} - \hat{w}_k, \frac{e_k}{\tau}\rangle + \langle \hat{w}_k - u, \xi_k\rangle \\
&\quad + \frac{1}{2\tau}\|\hat{u}_{k+1} - \hat{w}_k\|^2 + \frac{\varepsilon_{2,k}^2}{2\tau} \\
&= d_{\lambda_k}(u) + \langle \hat{w}_k - u, \xi_k\rangle - \frac{\tau}{2}\|\frac{\hat{u}_{k+1} - \hat{w}_k}{\tau}\|^2 + \tau\langle \frac{\hat{u}_{k+1} - \hat{w}_k}{\tau}, \frac{e_k}{\tau}\rangle + \frac{\varepsilon_{2,k}^2}{2\tau} \\
&= d_{\lambda_k}(u) + \langle \hat{w}_k - u, \xi_k\rangle - \frac{\tau}{2}\|\xi_k\|^2 + \frac{1}{2\tau}\left(\|e_k\|^2 + \varepsilon_{2,k}^2\right) \\
&\leqslant d_{\lambda_k}(u) + \langle \hat{w}_k - u, \xi_k\rangle - \frac{\tau}{2}\|\xi_k\|^2 + \frac{\varepsilon_{1,k}^2}{2\tau},
\end{aligned} \tag{5.4}$$

which can be seen as a noisy version of (4.11). We divide the rest of the proof in three steps. Since the former ones are analogous to the calculations done in the error-free case, we will skip some of the details.

Step 1 We show that for every $k \geqslant 1$, there holds:

$$\hat{\mathcal{E}}(k+1) - \hat{\mathcal{E}}(k) \leqslant t_{k+1}\Big(d_{\lambda_k}(u^\dagger) - \inf d_0\Big) + \frac{t_{k+1}}{\tau}\langle e_k, \hat{z}_k - u^\dagger\rangle + \frac{t_{k+1}^2}{2\tau}\varepsilon_{2,k}^2 \tag{5.5}$$

To prove this, we write the descent inequality (5.4) first for $u = \hat{u}_k$

$$d_{\lambda_k}(\hat{u}_{k+1}) \leqslant d_{\lambda_k}(\hat{u}_k) + \langle \hat{w}_k - \hat{u}_k, \xi_k\rangle - \frac{\tau}{2}\|\xi_k\|^2 + \frac{\varepsilon_{1,k}^2}{2\tau}, \tag{5.6}$$

and then for $u = u^\dagger$

$$d_{\lambda_k}(\hat{u}_{k+1}) \leqslant d_{\lambda_k}(u^\dagger) + \langle \hat{w}_k - u^\dagger, \xi_k\rangle - \frac{\tau}{2}\|\xi_k\|^2 + \frac{\varepsilon_{1,k}^2}{2\tau}. \tag{5.7}$$

We now multiply (5.6) by $t_{k+1}-1$ and add it to (5.7), thus getting:

$$t_{k+1}d_{\lambda_k}(\hat{u}_{k+1}) \leqslant (t_{k+1}-1)d_{\lambda_k}(\hat{u}_k) + d_{\lambda_k}(u^\dagger) \\ + \langle \xi_k, (t_{k+1}-1)(\hat{w}_k-\hat{u}_k)+\hat{w}_k-u^\dagger\rangle - \frac{t_{k+1}\tau}{2}\|\xi_k\|^2 + \frac{t_{k+1}}{2\tau}\varepsilon_{1,k}^2$$

We apply the property $(t_{k+1}-1)(\hat{w}_k-\hat{u}_k)+\hat{w}_k=\hat{z}_k$ (see (4.9)) and write (5.2) as

$$\begin{aligned} & t_{k+1}(d_{\lambda_k}(\hat{u}_{k+1})-d_{\lambda_k}(u^\dagger)) \\ \leqslant\ & (t_{k+1}-1)(d_{\lambda_k}(\hat{u}_k)-d_{\lambda_k}(u^\dagger)) + \frac{1}{2\tau t_{k+1}}\left(\|\hat{z}_k-u^\dagger\|^2 - \|\hat{z}_k-u^\dagger-\tau t_{k+1}\xi_k\|^2\right) + \frac{t_{k+1}}{2\tau}\varepsilon_{1,k}^2. \end{aligned}$$

From the identity $-\tau t_{k+1}\xi_k = \hat{z}_{k+1}-\hat{z}_k-t_{k+1}e_k$, we deduce:

$$\begin{aligned} & t_{k+1}(d_{\lambda_k}(\hat{u}_{k+1})-d_{\lambda_k}(u^\dagger)) + \frac{1}{2\tau t_{k+1}}\|\hat{z}_{k+1}-u^\dagger\|^2 \\ \leqslant\ & (t_{k+1}-1)(d_{\lambda_k}(\hat{u}_k)-d_{\lambda_k}(u^\dagger)) + \frac{1}{2\tau t_{k+1}}\|\hat{z}_k-u^\dagger\|^2 + \frac{1}{\tau}\langle \hat{z}_{k+1}-u^\dagger, e_k\rangle + \frac{t_{k+1}}{2\tau}\left(\varepsilon_{1,k}^2-\|e_k\|^2\right). \\ =\ & (t_{k+1}-1)(d_{\lambda_k}(\hat{u}_k)-d_{\lambda_k}(u^\dagger)) + \frac{1}{2\tau t_{k+1}}\|\hat{z}_k-u^\dagger\|^2 + \frac{1}{\tau}\langle \hat{z}_{k+1}-u^\dagger, e_k\rangle + \frac{t_{k+1}}{2\tau}\varepsilon_{2,k}^2. \end{aligned}$$

We now multiply everything by t_{k+1}, re-arrange and get

$$\begin{aligned} & t_{k+1}^2\Big(d_{\lambda_k}(\hat{u}_{k+1})-d_{\lambda_k}(u^\dagger)\Big) + \frac{1}{2\tau}\|\hat{z}_{k+1}-u^\dagger\|^2 \\ \leqslant\ & t_k^2\Big(d_{\lambda_k}(\hat{u}_k)-d_{\lambda_k}(u^\dagger)\Big) + (t_{k+1}^2-t_{k+1}-t_k^2)\Big(d_{\lambda_k}(\hat{u}_k)-d_{\lambda_k}(u^\dagger)\Big) + \frac{1}{2\tau}\|\hat{z}_k-u^\dagger\|^2 \\ & + \frac{t_{k+1}}{\tau}\langle e_k, \hat{z}_{k+1}-u^\dagger\rangle + \frac{t_{k+1}^2}{2\tau}\varepsilon_{2,k}^2. \end{aligned}$$

Using now that $d_{\lambda_k}(\hat{u}_k) \geqslant \inf d_0$ (see Lemma A.1(iv)), adding and subtracting in the parentheses the term $\inf d_0$ and after recalling the definition of $\mathcal{E}$ in (5.3), we get:

$$\hat{\mathcal{E}}(k+1) + (t_k^2+t_{k+1}-t_{k+1}^2)\Big(d_{\lambda_k}(u_k)-\inf d_0\Big) \\ \leqslant \hat{\mathcal{E}}(k) + t_{k+1}\Big(d_{\lambda_k}(u^\dagger)-\inf d_0\Big) + \frac{t_{k+1}}{\tau}\langle e_k, \hat{z}_k-u^\dagger\rangle + \frac{t_{k+1}^2}{2\tau}\varepsilon_{2,k}^2,$$

whence we deduce condition (5.5) since $t_k^2+t_{k+1}-t_{k+1}^2 \geqslant 0$ and $d_{\lambda_k}(u^\dagger)-\inf d_0 \geqslant 0$ (see (4.5)). Iterating recursively (5.5), Cauchy-Schwartz inequality yields

$$\hat{\mathcal{E}}(k) \leqslant \hat{\mathcal{E}}(1) + \sum_{j=1}^{k-1} t_{j+1}\Big(d_{\lambda_j}(u^\dagger)-\inf d_0\Big) + \sum_{j=1}^{k-1}\frac{t_{j+1}}{\tau}\varepsilon_{3,j}\|\hat{z}_{j+1}-u^\dagger\| + \sum_{j=1}^{k-1}\frac{t_{j+1}^2}{2\tau}\varepsilon_{2,j}^2, \tag{5.8}$$

which will be used to deduce the following stability estimate. Next, we study separately the sums appearing on the right-hand side of (5.8).

Step 2 For the first term in (5.8), following the proof of Theorem 4.6, we get

$$\sum_{j=1}^{k-1} t_{j+1}\Big(d_{\lambda_j}(u^\dagger)-\inf d_0\Big) \leqslant c\sum_{j=1}^{k-1} t_{j+1}\lambda_{\lambda_j}^{\frac{1}{q-1}} \leqslant c\Lambda < +\infty.$$

where c is defined in (3.4), and Λ is finite thanks to assumption (P_3).

Step 3 To bound the second sum in (5.8), we observe that by definition $\hat{\mathcal{E}}(k) \geqslant \frac{1}{2\tau}\|\hat{z}_k - u^\dagger\|^2$. Then, we set $C = 2\tau(\hat{\mathcal{E}}(1) + c\Lambda)$ and we derive

$$\|\hat{z}_k - u^\dagger\|^2 \leqslant C + \sum_{j=1}^{k-1} t_{j+1}^2 \varepsilon_{2,j}^2 + 2\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\|\hat{z}_{j+1} - u^\dagger\|.$$

Lemma A.5 applied to $a_k = \|\hat{z}_k - u^\dagger\|$, $b_k = 2t_{k+1}\varepsilon_{3,k}$, $c_{k-1} = C + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2$ implies

$$\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\|\hat{z}_{j+1} - u^\dagger\| \leqslant \Big(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\Big)\Big(\sqrt{C + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2} + 2\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\Big).$$

Combining altogether in (5.8), we deduce

$$\hat{\mathcal{E}}(k) \leqslant \frac{C}{2\tau} + \sum_{j=1}^{k-1} \frac{t_{j+1}^2}{2\tau}\varepsilon_{2,j}^2 + \frac{1}{\tau}\Big(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\Big)\Big(\sqrt{C + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2} + 2\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\Big). \tag{5.9}$$

Young's inequality applied to the product in the right hand side of (5.9) yields

$$\begin{aligned}
&\frac{1}{\tau}\Big(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\Big)\left(\sqrt{C + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2} + 2\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right)\\
=\ &\frac{1}{\tau}\Big(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\Big)\left(\sqrt{C + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2}\right) + \frac{2}{\tau}\Big(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\Big)^2\\
\leqslant\ &\frac{5}{2\tau}\left(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\right)^2 + \frac{C}{2\tau} + \frac{1}{2\tau}\sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{3,j}^2
\end{aligned}$$

From (5.9) we thus obtain

$$\hat{\mathcal{E}}(k) \leqslant \frac{C}{\tau} + \frac{1}{\tau}\sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2 + \frac{5}{2\tau}\Big(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\Big)^2. \tag{5.10}$$

To conclude, we use Lemma A.1(v) and deduce

$$\hat{\mathcal{E}}(k) \geqslant t_k^2(d_{\lambda_k}(\hat{u}_k) - \inf d_0) \geqslant t_k^2(d_0(\hat{u}_k) - \inf d_0) \geqslant \frac{t_k^2\sigma}{2}\|\hat{x}_k - x^\dagger\|^2,$$

that combined with (5.10) gives the desired stability estimate (5.2). □

5.3 Early stopping

We provide early stopping results guaranteeing the iterative regularization properties of ***(I3D)***. Different convergence rates are obtained depending on type of approximation considered (see Definition 5.3). The results follow from Theorem 5.5 considering assumption (E_i) for $i \in \{1, 2, 3\}$.

Theorem 5.6 (Early stopping for type 1 errors) *Assume that* (L_1)-(L_3), (R_1)-(R_2), (P_1)-(P_4) *hold true, and suppose that* $\lambda_k = \Theta(k^{-\theta})$ *with* $\theta > 2(q-1)$. *Let* $(\hat{x}_k)$ *be the sequence generated by* ***(I3D)*** *with noisy datum* $\hat{y}$, *and assume that* (E_i) *holds with* $i = 1$, $\varepsilon_{2,k} = O(\delta\lambda_k^{-r_2})$, $\varepsilon_{3,k} = O(\delta\lambda_k^{-r_3})$ *for some* $\delta > 0$ *and* $r_2, r_3 \geqslant 0$. *Set:*

$$\alpha := \max\left\{\frac{2}{3+2r_2\theta}, \frac{1}{2+r_3\theta}\right\}.$$

Then, any early stopping rule with $k(\delta) = \Theta(\delta^{-\alpha})$ *verifies:*

$$\|\hat{x}_{k(\delta)} - x^\dagger\| = O\left(\delta^\alpha\right), \quad \textit{for } \delta \searrow 0.$$

Proof. We apply estimate (5.2) from Theorem 5.5. After substituting the expression for $\varepsilon_{2,k}$ and $\varepsilon_{3,k}$, we get:

$$t_k^2\|\hat{x}_k - x^\dagger\|^2 = O\Big(1 + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2 + \Big(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\Big)^2\Big) = O(1 + \delta^2 k^{3+2r_2\theta} + \delta^2 k^{4+2r_3\theta}).$$

In correspondence with the stopping time $k(\delta)$, and using the fact that $t_{k(\delta)} = \Theta(k(\delta))$, we deduce from above:

$$\begin{aligned}\|\hat{x}_{k(\delta)} - x^\dagger\|^2 &= O\left(\delta^{2\alpha} + \delta^{2-\alpha(1+2r_2\theta)} + \delta^{2-2\alpha(1+r_3\theta)}\right)\\ &= O\left(\delta^{\min\{2\alpha; 2-\alpha(1+2r_2\theta), 2-2\alpha(1+r_3\theta)\}}\right).\end{aligned}$$

Let us now define $\beta := \min\{\frac{1}{2} + r_2\theta; 1 + r_3\theta\}$. We easily see that

$$\min\{2-\alpha(1+2r_2\theta); 2-2\alpha(1+r_3\theta)\} = 2 - 2\alpha\beta,$$

so that $\min\{2\alpha, 2-\alpha(1+2r_2\theta), 2-2\alpha(1+r_3\theta)\} = \min\{2\alpha, 2-2\alpha\beta\}$, which is maximal for $\alpha = \frac{1}{1+\beta}$. □ The analogous results for errors of type 2 and 3 are straightforward.

Theorem 5.7 (Early stopping for type 2 errors) *Assume that the assumptions* (L_1)-(L_3), (R_1)-(R_2), (P_1)-(P_4) *hold true, and suppose that* $\lambda_k = \Theta(k^{-\theta})$ *with* $\theta > 2(q-1)$. *Let* $(\hat{x}_k)$ *be the sequence generated by* ***(I3D)*** *with noisy datum* $\hat{y}$, *and assume that* (E_i) *holds with* $i = 2$, $\varepsilon_{2,k} = O(\delta\lambda_k^{-r_2})$ *for some* $\delta > 0$ *and* $r_2 \geqslant 0$. *Then, any early stopping rule with* $k(\delta) = \Theta(\delta^{-\frac{2}{3+2\theta r}})$ *verifies:*

$$\|\hat{x}_{k(\delta)} - x^\dagger\| = O\left(\delta^{\frac{2}{3+2\theta r}}\right), \quad \textit{for } \delta \searrow 0.$$

Proof. For type 2 approximation (5.2) $\varepsilon_{3,k} \equiv 0$, and we get

$$t_k^2\|\hat{x}_k - x^\dagger\|^2 = O\Big(1 + \sum_{j=1}^{k-1} t_{j+1}^2\varepsilon_{2,j}^2\Big) = O\Big(1 + \sum_{j=1}^{k-1} \delta^2 j^{2+2r\theta}\Big) = O(1 + \delta^2 k^{3+2r\theta}).$$

In correspondence with any stopping time $k(\delta) = \Theta(\delta^{-\alpha})$, we thus have:

$$\|\hat{x}_{k(\delta)} - x^\dagger\|^2 = O\left(k(\delta)^{-2} + \delta^2 k(\delta)^{1+2r\theta}\right) = O\left(\delta^{2\alpha} + \delta^{2-\alpha(1+2r\theta)}\right).$$

The term on the right-hand side is minimized when $\alpha = \frac{2}{3+2\theta r}$. □

Theorem 5.8 (Early stopping for type 3 errors) *Assume that the assumptions* (L_1)-(L_3), (R_1)-(R_2), (P_1)-(P_4) *hold true, and suppose that* $\lambda_k = \Theta(k^{-\theta})$ *with* $\theta > 2(q-1)$. *Let* $(\hat{x}_k)$ *be the sequence generated by* ***(I3D)*** *with noisy datum* $\hat{y}$, *and assume that* (E_i) *holds with* $i = 3$ *with* $\varepsilon_{3,k} = O(\delta \lambda_k^{-r_3})$ *for some* $\delta > 0$ *and* $r_3 \geqslant 0$. *Then, any early stopping rule with* $k(\delta) = \Theta(\delta^{-\frac{1}{2+\theta r}})$ *verifies:*

$$\|\hat{x}_{k(\delta)} - x^\dagger\| = O\left(\delta^{\frac{1}{2+\theta r}}\right), \quad \text{for } \delta \searrow 0.$$

Proof. Assuming type 3 errors means that in the estimate (5.2) $\varepsilon_{2,k} \equiv 0$, so that:

$$t_k^2 \|\hat{x}_k - x^\dagger\|^2 = O(1) + O\Big(\sum_{j=1}^{k-1} t_{j+1}\varepsilon_{3,j}\Big)^2 = O(1) + O\Big(\sum_{j=1}^{k-1} \delta j^{1+r\theta}\Big)^2 = O(1 + \delta^2 k^{4+2r\theta}).$$

In correspondence with the stopping time $k(\delta) = \Theta(\delta^{-\alpha})$, we thus deduce:

$$\|\hat{x}_{k(\delta)} - x^\dagger\|^2 = O\left(k(\delta)^{-2} + \delta^2 k(\delta)^{2+2r\theta}\right) = O\left(\delta^{2\alpha} + \delta^{2-2\alpha(1+r\theta)}\right).$$

The term on the right-hand side is minimal whenever $\alpha = \frac{1}{2+\theta r}$. □

6 Applications to specific data-fit terms

We next apply the results from Section 5.3 to relevant data-fit terms. The following definition is useful.

Definition 6.1 (δ-perturbation) For given $\bar{y},\ \hat{y} \in \mathcal{Y}$ and $\delta \in \mathbb{R}_{++}$, we say that $\hat{y}$ is a δ-perturbation of $\bar{y}$ according to ℓ if:

$$\ell_{\hat{y}}(\bar{y}) = \ell(\bar{y}, \hat{y}) \leqslant \delta^q,$$

where $q \in [1, +\infty)$ is the conditioning exponent appearing in (L_3).

We show that a δ-perturbation in the data corresponds to a proximal mapping of $\ell_{\hat{y}}^*$ approximating the corresponding proximal mapping of $\ell_{\bar{y}}^*$ in the sense of Definition 5.3 and with some precision $\varepsilon(\delta)$ depending on the noise level δ.

6.1 Additive data-fit terms

For additive data-fit functions (see Example 2.2), a δ-perturbation corresponds to a type 3 approximation of the proximal mapping.

Proposition 6.2 (Additive data-fit terms lead to type 3 errors) *Let* $\mathcal{N} \in \Gamma_0(\mathcal{Y})$ *and assume that* $\ell_{y_2}(y_1) = \mathcal{N}(y_2 - y_1)$, *for every* $(y_1, y_2) \in \mathcal{Y}^2$. *For given* $(\delta, \tau, \lambda) \in (0, +\infty)^3$, *let* $\hat{y} \in \mathbb{B}(\bar{y}, \varrho)$ *be a* δ-*perturbation of* $\bar{y}$ *in the sense of Definition 6.1. Then:*

$$(\forall z \in \mathcal{Y}) \quad \hat{p} = \operatorname{prox}_{\frac{\tau}{\lambda}\ell_{\hat{y}}^*(\lambda\cdot)}(z) \ \approx_3^{\varepsilon}\ \bar{p} = \operatorname{prox}_{\frac{\tau}{\lambda}\ell_{\bar{y}}^*(\lambda\cdot)}(z).$$

with precision $\varepsilon = \tau\delta(q/\gamma)^{1/q}$ *and where* $q \geqslant 1$ *and* $\gamma > 0$ *are the conditioning parameters appearing in assumption* (L_3).

Proof. We need to find $e \in \mathcal{Y}$ and $\varepsilon \geqslant 0$ such that $\|e\| \leqslant \varepsilon$ and:

$$\frac{z+e-\hat{p}}{\tau} \in \frac{1}{\lambda}\partial\ell^*_{\bar{y}}(\lambda\cdot)(\hat{p}). \tag{6.1}$$

Due to the special form of the data-fit we start noting that for any $u \in \mathcal{Y}$ we have

$$\ell^*_{\bar{y}}(u) = \mathcal{N}^*(u) + \langle \bar{y}, u\rangle,$$

and the same holds for $\ell^*_{\hat{y}}$. Then

$$\partial\ell^*_{\hat{y}}(\lambda\cdot)(\hat{p}) = \lambda\partial\ell^*_{\hat{y}}(\lambda\hat{p}) = \lambda\partial\Big(\mathcal{N}^* + \langle \hat{y}, \cdot\rangle\Big)(\lambda\hat{p}) = \lambda\partial\mathcal{N}^*(\lambda\hat{p}) + \lambda\hat{y}.$$

By definition of $\hat{p}$ we have $(z-\hat{p})/\tau \in (1/\lambda)\partial\ell^*_{\hat{y}}(\lambda\cdot)(\hat{p}) = \partial\mathcal{N}^*(\lambda\hat{p}) + \hat{y}$, which, by simple algebraic manipulations, entails the required condition (6.1), since:

$$\frac{z-\hat{p}}{\tau} \in \partial\mathcal{N}^*(\lambda\hat{p}) + \bar{y} + (\hat{y}-\bar{y}) \iff \frac{z-\hat{p}+\tau(\bar{y}-\hat{y})}{\tau} \in \partial\mathcal{N}^*(\lambda\hat{p}) + \bar{y} = \frac{1}{\lambda}\partial\ell^*_{\bar{y}}(\lambda\cdot)(\hat{p}).$$

By setting $e = \tau(\bar{y}-\hat{y})$, we can find the required ε combining the local q-conditioning of the function $\ell_{\bar{y}}$ on $\mathbb{B}(\bar{y}, \varrho)$ assumed in (L_3) with the δ-perturbation assumption:

$$\|e\| = \tau\|\bar{y}-\hat{y}\| \leqslant \tau\left(\frac{q}{\gamma}\ell(\hat{y},\bar{y})\right)^{1/q} \leqslant \tau\left(\frac{q}{\gamma}\right)^{1/q}\delta =: \varepsilon,$$

where $\gamma > 0$ and $q \geqslant 1$ are the conditioning parameters. We can thus conclude that $\hat{p}$ is a ε-approximation of $\bar{p}$ with precision ε, as required. □

Thanks to Proposition 6.2, we can derive early-stopping result by applying Theorem 5.8 to additive data-fit terms with the above choice of ε.

Corollary 6.3 (Early stopping for additive data-fit terms) *Let $\mathcal{N} \in \Gamma_0(\mathcal{Y})$ and set $\ell_{y_2}(y_1) = \mathcal{N}(y_2 - y_1)$, for every $(y_1, y_2) \in \mathcal{Y}^2$ Assume that the assumptions (L_1)-(L_3), (R_1)-(R_2), (P_1)-(P_4) hold, and that $\lambda_k = \Theta(k^{-\theta})$ with $\theta > 2(q-1)$. Let $(\hat{x}_k)$ be the sequence generated by **(I3D)** with $\hat{y} \in \mathbb{B}(\bar{y}, \varrho)$, such that $\hat{y}$ is a δ-perturbation of $\bar{y}$. Then, any early stopping rule with $k(\delta) = \Theta(\delta^{-1/2})$ verifies:*

$$\|\hat{x}_{k(\delta)} - x^\dagger\| = O(\delta^{\frac{1}{2}}), \quad \text{for } \delta \searrow 0. \tag{6.2}$$

Remark 6.4 (Optimality of the rates) The convergence rate in (6.2) is optimal for regularization methods for additive data-fit terms [38]. Among inertial algorithms, optimal convergence rates for different regularizers but only quadratic data-fit terms have been proved in [48, 45]. For more general additive data-fits (e.g. the ℓ^1-norm, see Example 2.2), in [24] the authors prove a rate $O(\delta^{1/2})$ on the Bregman distance, which is different from (6.2). To our knowledge, our result is the first showing optimal convergence rates for iterative regularization methods for general data-fit term improving the estimates in [39] that showed a rate $O(\delta^{1/3})$.

Remark 6.5 (Different growth for t_k) As noted in Remark 4.8, if we replace $t_k = \Theta(t_k)$ by $t_k = \Theta(k^\beta)$, then $\beta \leqslant 1$, and $\beta = 1$ gives the fastest convergence rate for true datumum $\bar{y}$. Corollary 6.3 implies that also for noisy data $\hat{y}$, any stopping rule with $k(\delta) = \Theta(\delta^{-1/(1+\beta)})$ verifies $\|\hat{x}_{k(\delta)} - x^\dagger\| = O(\delta^{\frac{\beta}{\beta+1}})$ for $\delta \searrow 0$, where again the best rate is achieved for $\beta = 1$.

6.2 KL divergence

We consider the Kullback-Leibler (KL) divergence as an example of non-additive data-fit term. KL divergence is often used to model data corrupted by Poisson noise. We show that the KL divergence δ-perturbations lead to type 2 approximations. We recall that the KL divergence is locally 2-conditionned (see Example 2.2).

Proposition 6.6 *Assume that, $\ell_{y_2}(y_1) = \mathrm{KL}(y_2; y_1)$ for every $(y_1, y_2) \in \mathcal{Y}^2$. For $(\delta, \tau, \lambda) \in (0, +\infty)^3$, let $\hat{y} \in \mathbb{B}(\bar{y}, \varrho)$ be a δ-perturbation of $\bar{y}$. Then*

$$(\forall z \in \mathcal{Y}) \quad \hat{p} = \mathrm{prox}_{\frac{\tau}{\lambda}\ell^*_{\hat{y}}(\lambda\cdot)}(z) \;\approx_2^{\varepsilon}\; \bar{p} = \mathrm{prox}_{\frac{\tau}{\lambda}\ell^*_{\bar{y}}(\lambda\cdot)}(z).$$

with $\varepsilon = \sqrt{2\tau}\delta/\lambda$.

Proof. It is enough to prove that for all $z \in \mathcal{Y}$

$$\frac{\lambda(z-\hat{p})}{\tau} \in \partial_{\frac{\lambda\varepsilon^2}{2\tau}} \mathrm{KL}^*_{\bar{y}}(\lambda\cdot)(\hat{p}) = \lambda\partial_{\frac{\lambda\varepsilon^2}{2\tau}} \mathrm{KL}^*_{\bar{y}}(\lambda\hat{p}), \quad \Longleftrightarrow \quad \frac{z-\hat{p}}{\tau} \in \partial_{\frac{\lambda\varepsilon^2}{2\tau}} \mathrm{KL}^*_{\bar{y}}(\lambda\hat{p}).$$

We set $x = (z-\hat{p})/\tau \in \mathcal{Y}$ and consider the function $g : \mathcal{Y} \to \mathbb{R}^d \cup \{+\infty\}$ defined by

$$g(w) = \frac{\mathrm{KL}_{\bar{y}}}{\lambda}(w), \quad \text{for all } w \in \mathcal{Y}. \tag{6.3}$$

By standard property of convex conjugates we have that for any $u \in \mathcal{Y}$

$$g^*(u) = \left(\frac{\mathrm{KL}_{\bar{y}}}{\lambda}\right)^*(u) = \frac{1}{\lambda}\mathrm{KL}^*_{\bar{y}}(\lambda u). \tag{6.4}$$

We now claim that $x \in \partial_{\frac{\lambda\varepsilon^2}{2\tau}} g^*(\hat{p})$. To show that, we apply the Young-Fenchel inequality of Lemma A.6 to g with $x^* = \hat{p}$. Our objective is thus to show that:

$$g(x) + g^*(\hat{p}) \leqslant \langle x, \hat{p}\rangle + \frac{\lambda\varepsilon^2}{2\tau},$$

which, by definitions (6.3) and (6.4) and upon multiplication by λ, coincides with:

$$\mathrm{KL}_{\bar{y}}(x) + \mathrm{KL}^*_{\bar{y}}(\lambda\hat{p}) \leqslant \langle x, \lambda\hat{p}\rangle + \frac{\lambda^2\varepsilon^2}{2\tau}. \tag{6.5}$$

Using expressions (**??**) and (A.1) for KL and its convex conjugate, we express the sum on the left hand side of (6.5) as:

$$\mathrm{KL}_{\bar{y}}(x) + \mathrm{KL}^*_{\bar{y}}(\lambda\hat{p}) = \sum_{i=1}^{d} \Big(\bar{y}_i \log\frac{\bar{y}_i}{x_i} - \bar{y}_i + x_i - \bar{y}_i\log(1-\lambda\hat{p}_i)\Big). \tag{6.6}$$

Furthermore, by definition of $\hat{p}$, we have that component-wise there holds:

$$\frac{\lambda}{\tau}(z_i - \hat{p}_i) \in \lambda\partial\mathrm{kl}^*_{\hat{y}_i}(\lambda\hat{p}_i) \quad \Longleftrightarrow \quad x_i \in \partial\mathrm{kl}^*_{\hat{y}_i}(\lambda\hat{p}_i),$$

which, since $\mathrm{kl}^*_{\hat{y}_i}$ is differentiable (see formula (A.1)), entails that for every $i = 1, \dots, d$ the element x_i can be written as $x_i = \hat{y}_i / 1 - \lambda \hat{p}_i$. Substitute this expression in the formula (6.6) to derive

$$\begin{aligned}
\mathrm{KL}_{\bar{y}}(x) + \mathrm{KL}^*_{\bar{y}}(\lambda \hat{p}) &= \sum_{i=1}^{d} \underbrace{\bar{y}_i \log \bar{y}_i - \bar{y}_i \log \hat{y}_i - \bar{y}_i + \hat{y}_i}_{\mathrm{kl}(\bar{y}_i; \hat{y}_i)} \\
&\quad + \cancel{\bar{y}_i \log(1 - \lambda \hat{p}_i)} + \underbrace{\big(\hat{y}_i / (1 - \lambda \hat{p}_i)\big)}_{x_i} \lambda \hat{p}_i - \cancel{\bar{y}_i \log(1 - \lambda \hat{p}_i)} \\
&= \mathrm{KL}_{\bar{y}}(\hat{y}) + \langle x, \lambda \hat{p} \rangle \\
&\leqslant \delta^2 + \langle x, \lambda \hat{p} \rangle,
\end{aligned}$$

where the last inequality follows from the perturbation assumption $\mathrm{KL}_{\bar{y}}(\hat{y}) \leqslant \delta^2$. We thus get (6.5) by choosing $\varepsilon = \sqrt{2\tau}\delta/\lambda$, which concludes the proof. □ From Proposition 6.6 and Theorem 5.7, we derive stopping rules for the KL divergence.

Corollary 6.7 (Early stopping for Kullback-Leibler divergence) *Let $\ell_{y_2}(y_1) = \mathrm{KL}(y_2; y_1)$ for every $(y_1, y_2) \in \mathcal{Y}^2$. Assume that the assumptions (L_1)-(L_3), (R_1)-(R_2), (P_1)-(P_4) hold true, and suppose that $\lambda_k = \Theta(k^{-\theta})$ with $\theta > 2$. Let $(\hat{x}_k)$ be the sequence generated by **(I3D)** given $\hat{y}$, such that $\hat{y}$ is a δ-perturbation of $\bar{y}$ in the sense of Definition 6.1. Then, any early stopping rule with $k(\delta) = \Theta(\delta^{-\frac{2}{3+2\theta}})$ verifies*

$$\|\hat{x}_{k(\delta)} - x^\dagger\| = O(\delta^{\frac{2}{3+2\theta}}), \quad \text{for } \delta \searrow 0. \tag{6.7}$$

Remark 6.8 It is hard to assess the quality of the rate in (6.7) since the the notion of optimality in [38] only applies to additive noise. In the context of Bregman divergences, a similar analysis has been pursued in [24, Section 4.2, estimate (4.3)]. The estimate in [24] leads to a rate of order $\delta^{1/4}$ for suitable choices of the regularization parameter. In comparison, our estimate (6.7) is sharper and more explicit. As for additive data-fit terms, the use of inertia improves the rates in [39].

Remark 6.9 (The Kullback-Leibler divergence does not lead to type 3 errors) The convergence rates for additive data-fit terms proved in Corollary 6.3 are better than the rate for the KL divergence, due to the fact that for the KL divergence we proved that δ-perturbations correspond to type 2 errors, instead of type 3 errors. Indeed, Lemma A.3 in the Appendix shows that the error in the evaluation of proximal points for the KL divergence can not be cast in a type 3 approximation.

7 Conclusions and outlook

In this paper we proposed an inertial dual diagonal method to solve inverse problems for a wide class of data-fit and regularization terms, possibly corrupted by noise. On the one hand we established convergence results both for continuous and discrete dynamics. On the other hand we derived stability results and corresponding stopping rules, characterizing the regularization properties of the proposed method. A number of open questions are left for future study. It would be interesting to consider wider class of problems for example allowing for regularization terms that are convex but not strongly convex, and possibly non convex data fidelity terms. From an algorithmic point of view, it would be interesting to consider alternative approaches, for example considering stochastic methods. Finally, it would be interesting to investigate the numerical properties of the proposed method for practical problems.

A Auxiliary results

We gather some relevant results.

A.1 Properties of the dual diagonal function

We first consider $\mathcal{R}_A$, ℓ_y^* defined in (3.1) and on the diagonal dual function d_λ and its limit d_0 defined in **(D_λ)** and **(D_0)**, respectively. For similar results see also [39].

Lemma A.1 *Under the assumptions (L_1)-(L_3) and (R_1)-(R_2), we have:*

(i) *$\mathcal{R}_A$ is differentiable and ∇R_A^* is Lipschitz continuous, with Lipschitz constant equal to $\sigma^{-1}\|A\|^2$.*

(ii) *$\forall y \in \mathcal{Y}$, $\ell_y^*(0) = 0$ and $\partial \ell_y^*(0) = \{y\}$.*

(iii) *There holds: $\operatorname{argmin} d_0 \neq \emptyset$.*

(iv) *$\forall u \in \mathcal{Y}$, the function $\lambda \in [0, +\infty) \mapsto d_\lambda(u)$ is nondecreasing.*

(v) *$\forall t > 0, \forall u \in \mathcal{Y}$, if $x := \nabla R^*(-A^*u)$, then $\frac{\sigma}{2}\|x - x^\dagger\|^2 \leqslant d_0(u) - \inf d_0$.*

(vi) *$\forall u^\dagger \in \operatorname{argmin} d_0$, if $\lambda\|u^\dagger\| \leqslant \frac{\gamma}{q}\varrho^{q-1}$, then*

$$d_\lambda(u^\dagger) - \inf d_0 \leqslant \begin{cases} 0 & \text{if } q = 1, \\ (1 - \frac{1}{q})\gamma^{-1/(q-1)}\|u^\dagger\|^{q/(q-1)}\lambda^{1/(q-1)} & \text{if } q > 1. \end{cases}$$

Proof. (i): follows from the strong convexity of R, see, e.g., [21, Theorem 18.15].
(ii): it is a simple consequence of the properties of the Fenchel transform as it can be found, e.g., in [21, Proposition 13.10(i) & Corollary 16.30].
(iii) and (v): follow from [39, Lemma 5] by simply taking $f = R$ and $g = \delta_{\{\bar{y}\}}$, while property (iv) has been proved in [39, Proposition 2(i)].
(vi): it is enough to verify that $\ell_{\bar{y}}(\cdot)$ is q-well conditioned in the sense of [39, Definition 1], while assumption (L_3) holds only locally. To check this, we introduce the function $\psi : \mathbb{R} \to \mathbb{R}$ defined for the $\varrho > 0$ appearing in (L_3) by:

$$\psi t \mapsto \begin{cases} \frac{\gamma}{q}|t|^q & \text{if } |t| \leqslant \varrho, \\ \frac{\gamma}{q}\varrho^{q-1}|t| & \text{if } |t| > \varrho. \end{cases}$$

From (L_3), we easily deduce that $\ell_{\bar{y}}(y) \geqslant \psi(\|y - \bar{y}\|)$ for all $y \in \mathcal{Y}$ (see [57, Corollary 3.4.2]). Note that ψ is not convex for $q > 1$, so in this case we consider instead the function

$$m : \mathbb{R} \to \mathbb{R}, \quad t \mapsto \begin{cases} \frac{\gamma}{q}|t|^q & \text{if } |t| \leqslant q^{1/(1-q)}\varrho, \\ \frac{\gamma}{q}\varrho^{q-1}|t| - \frac{\gamma}{q^{\frac{q}{q-1}}}\varrho^q(1 - \frac{1}{q}) & \text{if } |t| > q^{1/(1-q)}\varrho, \end{cases}$$

and define $m := \psi$ for $q = 1$. It is an easy exercise to verify that m is indeed a convex function on $\mathbb{R}$, and that $m(w) \leqslant \psi(w)$ for all $w \in \mathbb{R}$. Now, we can make use of [39, Lemma 2], which tells us that $d_\lambda(u) - \inf d_0 \leqslant \lambda^{-1} m^*(\|u\|\lambda)$. The desired result now follows from the computation of the

Fenchel transform of m. If $q = 1$, we have that $m(t) = \gamma|t|$, so classic Fenchel calculus entails that m^* is $\delta_{[-\gamma,\gamma]}$, the indicator function of $[-\gamma, \gamma]$. If $q > 1$, easy computations show that m^* reads

$$m^* : \mathbb{R} \to \mathbb{R}, \quad s \mapsto \begin{cases} (1 - \frac{1}{q})\gamma^{-1/(q-1)}|s|^{\frac{q}{q-1}} & \text{if } |s| \leqslant \frac{\gamma}{q}\varrho^{q-1}, \\ +\infty & \text{if } |s| > \frac{\gamma}{q}\varrho^{q-1}. \end{cases}$$

By now applying [39, Lemma 2] we conclude. □

A.2 Useful tools for KL computations

In this section, we report some computations and properties concerning the KL divergence defined in (2.2). For any $(u, y) \in (\mathbb{R}^d)^2$ we define $\mathrm{KL}(y, u)$ as in (2.2). Consider now the functions KL and kl with respect to the first argument only, and define $\mathrm{KL}_y(u) := \mathrm{KL}(y; u)$ and, similarly, its i-th component $\mathrm{kl}_{y_i}(u_i)$ for a fixed $y \in \mathbb{R}^d$. The component-wise expression for $\mathrm{KL}^*_y(w) = \sum_{i=1}^d \mathrm{kl}^*_{y_i}(w_i)$ can be then simply found by Fenchel calculus. It reads:

$$\mathrm{kl}^*_{y_i}(w_i) = \begin{cases} -y_i \log(1 - w_i) & \text{if } 1 - w_i > 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{A.1}$$

Proximal maps For every $i = 1, \ldots, d$, straightforward calculations show that

$$\mathrm{prox}_{\frac{\tau}{\lambda}\mathrm{kl}_{y_i}}(u_i) = \frac{1}{2}\Big(u_i - \frac{\tau}{\lambda} + \sqrt{\Big(u_i - \frac{\tau}{\lambda}\Big)^2 + 4\frac{\tau}{\lambda}y_i}\Big).$$

Furthermore, by applying Moreau's identity we have:

$$\mathrm{prox}_{\frac{\tau}{\lambda}\mathrm{kl}^*_{y_i}(\lambda\cdot)}(w_i) = \frac{1}{2\lambda}\Big((1 + \lambda w_i) - \sqrt{(1 - \lambda w_i)^2 + 4\lambda\tau y_i}\Big). \tag{A.2}$$

The following lemma implies the q-conditioning of the Kullback-Leibler divergence.

Lemma A.2 (2-conditioning of the KL data-fit term) *Let $\bar{y} \in \mathbb{R}^d$ and $\varrho \in (0, +\infty)$. Then,*

$$(\forall y \in \mathbb{B}(\bar{y}, \varrho)) \quad \mathrm{KL}(\bar{y}, y) \geqslant \left(\frac{1}{\varrho c^2} + \frac{1}{\varrho^2 c}\ln\frac{c}{\varrho + c}\right)\|y - \bar{y}\|^2, \text{ where } c = d\|\bar{y}\|_\infty.$$

Proof. Let $y \in \mathbb{B}(\bar{y}, \varrho)$. By [39, Lemma 10.2], we have that

$$\mathrm{KL}(\bar{y}, y) \geqslant cm(\|y - \bar{y}\|), \text{ where } m(t) = c^{-1}|t| - \ln\left(1 + c^{-1}|t|\right). \tag{A.3}$$

To get the desired result, we need to find a quadratic lower bound for m over $[-\varrho, \varrho]$. For simplicity, let us consider the change of variable $s = c^{-1}|t| \in [0, c^{-1}\varrho]$. Since the statement is trivially valid for $y = \bar{y}$, we can assume that $s > 0$ and write

$$s - \ln(1 + s) = s^2\phi(s), \text{ where } \phi(s) := \frac{s - \ln(1 + s)}{s^2}.$$

To conclude, we only need to verify that ϕ is decreasing on $]0, +\infty[$. Indeed, this would imply that $m(t) \geqslant c^{-2}t^2\phi(c^{-1}\varrho)$, which together with (A.3) would complete the proof. To see that ϕ is

decreasing, we compute explicitly its derivative on $]0,+\infty[$ and see that $\phi'(s) \leqslant 0$ if and only if $\psi(s) := s(s+2) - 2(1+s)\ln(1+s) \geqslant 0$. Combining this with the fact that $\psi(0) = 0$, and that $\psi'(s) = 2(s - \ln(1+s))$ is positive $]0,+\infty[$ we conclude the proof. □

In the following we deal with proximal points of the dual of the KL divergence, corresponding to noise-free and noisy data $\bar{y}$ and $\hat{y}$, respectively. As shown in Proposition 6.6 a type 2 approximation in the sense of Definition 5.3 holds. The following proposition is a one-dimensional counterexample showing that a type 3 approximation – for which better convergence rates can be obtained – does not hold.

Proposition A.3 *Let $w \in \mathbb{R}$ and $\bar{y}, \hat{y} \in]0,+\infty[$. If $\operatorname{prox}_{\mathrm{kl}^*_{\hat{y}}}(w) \approx_3^{\varepsilon} \operatorname{prox}_{\mathrm{kl}^*_{\bar{y}}}(w)$ holds in the sense of Definition 5.3 for some $\varepsilon > 0$, then*

$$\varepsilon \geqslant \frac{2|\hat{y} - \bar{y}|}{(1-w) + \sqrt{(1-w)^2 + 4\hat{y}}}.$$

In particular, $\varepsilon \to +\infty$ when $w \to +\infty$.

Proof. Let $\varepsilon \geqslant 0$ such that the type 3 approximation property holds. By Definition 5.3, there exists $e \in \mathbb{R}$ such that $|e| \leqslant \varepsilon$ and $\operatorname{prox}_{\mathrm{kl}^*_{\hat{y}}}(w) = \operatorname{prox}_{\mathrm{kl}^*_{\bar{y}}}(w+e)$. Using the formula (A.2), we see that this is equivalent to

$$\frac{1}{2}\left[(1+w) - \sqrt{(1-w)^2 + 4\hat{y}}\right] = \frac{1}{2}\left[(1+w+e) - \sqrt{(1-w-e)^2 + 4\bar{y}}\right].$$

and we complete the proof by noting that the above equality is equivalent to

$$e\frac{1}{2}\left[(1-w) + \sqrt{(1-w)^2 + 4\hat{y}}\right] = \bar{y} - \hat{y}.$$

□

A.3 Miscellaneous

We here recall some technical lemmas which are used in several sections of the manuscript. The following Lemma is useful to characterize the speed of decay of the diagonal term $\lambda(\cdot)$ in assumption (Λ), see also Remark 3.2.

Lemma A.4 *Let $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ a decreasing function such that $\int_{\mathbb{R}_+} |\lambda(t)|^{1/2}\, dt < +\infty$. Then, the function $t \mapsto t\lambda(t)$ is integrable on $\mathbb{R}_+$.*

Proof. We first show that the function $t \mapsto t\sqrt{\lambda(t)}$ tends to zero as $t \to +\infty$. We have that for every $T > 0$:

$$\int_{T/2}^{+\infty} \sqrt{\lambda(t)}\, dt \geqslant \int_{T/2}^{T} \sqrt{\lambda(t)}\, dt \geqslant \frac{T}{2}\sqrt{\lambda(T)},$$

where the last inequality follows from the decreasing property of λ in the interval $[T/2, T]$. By taking limits, we get the required property:

$$\limsup_{T\to+\infty} \frac{T}{2}\sqrt{\lambda(T)} \leqslant \lim_{T\to+\infty} \int_{T/2}^{+\infty} \sqrt{\lambda(t)}\, dt = 0.$$

Now, from the observation

$$\lim_{t\to+\infty} \frac{t\lambda(t)}{\sqrt{\lambda(t)}} = \lim_{t\to+\infty} t\sqrt{\lambda(t)} = 0,$$

we deduce that there exists some $\overline{T} > 0$ such that $t\lambda(t) \leqslant \sqrt{\lambda(t)}$ for all $t \geqslant \overline{T}$. By thus taking $T > \overline{T}$, we have:

$$\int_0^T t\lambda(t)\,dt = \int_0^{\overline{T}} t\lambda(t)\,dt + \int_{\overline{T}}^T t\lambda(t)\,dt \leqslant \int_0^{\overline{T}} t\lambda(t)\,dt + \int_{\overline{T}}^T \sqrt{\lambda(t)}\,dt,$$

which by taking the supremum over all $T > \bar{T}$ on both sides entails:

$$\int_{\mathbb{R}_+} t\lambda(t)\,dt \leqslant \int_0^{\overline{T}} t\lambda(t)\,dt + \int_{\overline{T}}^{+\infty} \sqrt{\lambda(t)}\,dt < +\infty.$$

□

Next, we state and prove a variant of [7, Lemma 5.14] which we have used in the proof of Theorem 5.2 to get the final stability estimate (5.2).

Lemma A.5 *Let $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$ and $(c_k)_{k\in\mathbb{N}}$ be positive sequences, and assume that c_k is increasing. If*

$$(\forall k \in \mathbb{N}) \quad a_k^2 \leqslant c_k + \sum_{j=1}^{k-1} b_j a_{j+1},$$

then $\max_{j=1,\dots,k} a_j \leqslant \sqrt{c_k} + \sum_{j=1}^{k-1} b_j$, for every $k \in \mathbb{N}$.

Proof. Take $k \in \mathbb{N}$, and let $A_k := \max_{m=1,\dots,k} a_m$. Then, for all $1 \leqslant m \leqslant k$:

$$a_m^2 \leqslant c_m + \sum_{j=1}^{m-1} b_j a_{j+1} \leqslant c_k + A_k \sum_{j=1}^{k-1} b_j,$$

because c_k is increasing and b_j is positive. Therefore $A_k^2 \leqslant c_k + A_k \sum_{j=1}^{k-1} b_j$. Define $S_k = \sum_{j=1}^{k-1} b_j$. By computing and bounding the solutions of the previous inequality we conclude that

$$A_k \leqslant \frac{S_k + \sqrt{S_k + 4c_k}}{2} \leqslant S_k + \sqrt{c_k}.$$

□

We recall a useful characterisation of the elements in the ε-subdifferential of a function in $\Gamma_0(\mathcal{H})$. This property is used to prove Proposition 6.6, see also [57].

Lemma A.6 (Theorem 2.4.2, [57]) *Let $\mathcal{H}$ be an Hilbert space, let $f \in \Gamma_0(\mathcal{H})$, let $(x, u) \in \mathcal{H}^2$, and let $\varepsilon > 0$. Then, the following statements are equivalent:*

i) $u \in \partial_\varepsilon f(x)$;

ii) The following ε-Young-Fenchel inequality holds:

$$f(x) + f^*(u) \leqslant \langle u, x\rangle + \varepsilon;$$

iii) $x \in \partial_\varepsilon f^*(u)$.

References

[1] F. Alvarez. On the minimizing property of a second order dissipative system in Hilbert spaces. SIAM Journal on Control and Optimization, 38(4):1102–1119, 2000.

[2] V. Apidopoulos, J.-F. Aujol, and C. Dossal. Convergence rate of inertial forward–backward algorithm beyond nesterov's rule. Mathematical Programming, Nov 2018.

[3] V. Apidopoulos, J.-F. Aujol, and C. Dossal. The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case b $\leqslant 3$. SIAM Journal on Optimization, 28(1):551–574, 2018.

[4] H. Attouch. Viscosity solutions of minimization problems. SIAM Journal on Optimization, 6(3):769–806, 1996.

[5] H. Attouch, A. Cabot, Z. Chbani, and H. Riahi. Inertial forward–backward algorithms with perturbations: Application to Tikhonov regularization. Journal of Optimization Theory and Applications, 179(1):1–36, 2018.

[6] H. Attouch, A. Cabot, and M.-O. Czarnecki. Asymptotic behavior of nonautonomous monotone and subgradient evolution equations. Transactions of the American Mathematical Society, 370(2):755–790, 2018.

[7] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. Mathematical Programming, 168(1):123–175, 2018.

[8] H. Attouch, Z. Chbani, and H. Riahi. Combining fast inertial dynamics for convex optimization with Tikhonov regularization. Journal of Mathematical Analysis and Applications, 457(2):1065–1094, 2018.

[9] H. Attouch, Z. Chbani, and H. Riahi. Rate of convergence of the Nesterov accelerated gradient method in the subcritical case 3. ESAIM: COCV, 25:2, 2019.

[10] H. Attouch and R. Cominetti. A Dynamical Approach to Convex Minimization Coupling Approximation with the Steepest Descent Method. Journal of Differential Equations, 128(2):519–540, 1996.

[11] H. Attouch, M. Czarnecki, and J. Peypouquet. Coupling Forward-Backward with Penalty Schemes and Parallel Splitting for Constrained Variational Inequalities. SIAM Journal on Optimization, 21(4):1251–1274, 2011.

[12] H. Attouch and M.-O. Czarnecki. Asymptotic behavior of coupled dynamical systems with multiscale aspects. Journal of Differential Equations, 248(6):1315–1344, 2010.

[13] H. Attouch and M.-O. Czarnecki. Asymptotic behavior of gradient-like dynamical systems involving inertia and multiscale aspects. Journal of Differential Equations, 262(3):2745–2770, 2017.

[14] H. Attouch and J. Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. SIAM Journal on Optimization, 26(3):1824–1834, 2016.

[15] H. Attouch, J. Peypouquet, and P. Redont. A dynamical approach to an inertial forward-backward algorithm for convex minimization. SIAM Journal on Optimization, 24(1):232–256, 2014.

[16] J. Aujol and C. Dossal. Stability of over-relaxations for the forward-backward algorithm, application to fista. SIAM Journal on Optimization, 25(4):2408–2433, 2015.

[17] F. R. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In Advances in Neural Information Processing Systems, pages 105–112, 2009.

[18] M. Bachmayr and M. Burger. Iterative total variation schemes for nonlinear inverse problems. Inverse Problems, 25(10):105004, 2009.

[19] M. A. Bahraoui and B. Lemaire. Convergence of diagonally stationary sequences in convex optimization. Set-Valued Analysis, 2(1-2):49–61, 1994.

[20] A. B. Bakushinsky and M. Y. Kokurin. Iterative Methods for Approximate Solution of Inverse Problems, volume 577. Springer Science & Business Media, 2005.

[21] H. Bauschke and P. Combettes. Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics. Springer, 2017.

[22] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters, 31(3):167 – 175, 2003.

[23] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. SIAM Journal on Imaging Sciences, 2(1):183–202, 2009.

[24] M. Benning and M. Burger. Error estimates for general fidelities. Electron. Trans. Numer. Anal., 38:44–68, 2011.

[25] M. Benning and M. Burger. Modern regularization methods for inverse problems. Acta Numerica, 27:1–111, 2018.

[26] R. I. Boţ and T. Hein. Iterative regularization with a general penalty term—theory and application to L1 and TV regularization. Inverse Problems, 28(10):104010, 2012.

[27] K. Bredies, K. Kunisch, and T. Pock. Total Generalized Variation. SIAM Journal on Imaging Sciences, 3(3):492–526, 2010.

[28] P. Brianzi, F. Di Benedetto, and C. Estatico. Preconditioned iterative regularization in banach spaces. Computational Optimization and Applications, 54(2):263–282, 2013.

[29] M. Burger and S. Osher. A Guide to the TV Zoo. In Level Set and PDE Based Reconstruction Methods in Imaging, Lecture Notes in Mathematics, pages 1–70. Springer International Publishing, 2013.

[30] M. Burger, E. Resmerita, and L. He. Error estimation for Bregman iterations and inverse scale space methods in image restoration. Computing, 81(2-3):109–135, 2007.

[31] A. Cabot and L. Paoli. Asymptotics for some vibro-impact problems with a linear dissipation term. Journal de mathématiques pures et appliquées, 87(3):291–323, 2007.

[32] L. Calatroni, J. C. De Los Reyes, and C.-B. Schönlieb. Infimal convolution of data discrepancies for mixed noise removal. SIAM Journal on Imaging Sciences, 10(3):1196–1233, 2017.

[33] A. Chambolle and C. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". Journal of Optimization Theory and Applications, 166(3):968–982, Sep 2015.

[34] A. Chambolle and P.-L. Lions. Image recovery via total variation minimization and related problems. Numerische Mathematik, 76(2):167–188, 1997.

[35] M.-O. Czarnecki, N. Noun, and J. Peypouquet. Splitting Forward-Backward Penalty Scheme for Constrained Variational Problems. Journal of Convex Analysis, 23(2):531–565, 2016.

[36] C.-A. Deledalle, S. Vaiter, J. M. Fadili, and G. Peyré. Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection. SIAM Journal on Imaging Sciences, 7(4):2448–2487, 2014.

[37] O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. Mathematical Programming, 146(1):37–75, Aug 2014.

[38] H. W. Engl, M. Hanke, and A. Neubauer. Regularization of Inverse Problems, volume 375. Springer Science & Business Media, 1996.

[39] G. Garrigos, L. Rosasco, and S. Villa. Iterative regularization via dual diagonal descent. Journal of Mathematical Imaging and Vision, 60(2):189–215, 2018.

[40] M. Hintermüller and A. Langer. Subspace Correction Methods for a Class of Nonsmooth and Nonadditive Convex Variational Problems with Mixed l^1/l^2 Data-Fidelity in Image Processing. SIAM Journal on Imaging Sciences, 6(4):2134–2173, 2013.

[41] B. Kaltenbacher, A. Neubauer, and O. Scherzer. Iterative Regularization Methods for Nonlinear Ill-Posed Problems. De Gruyter, Berlin, Boston, 2008.

[42] B. Kaltenbacher, F. Schöpfer, and T. Schuster. Iterative methods for nonlinear ill-posed problems in banach spaces: convergence and applications to parameter identification problems. Inverse Problems, 25(6):065003, apr 2009.

[43] B. Kaltenbacher and I. Tomba. Convergence rates for an iteratively regularized Newton–Landweber iteration in banach space. Inverse Problems, 29(2):025010, jan 2013.

[44] W. Krichene, A. Bayen, and P. L. Bartlett. Accelerated Mirror Descent in Continuous and Discrete Time. In Advances in Neural Information Processing Systems, pages 2827–2835, 2015.

[45] S. Matet, L. Rosasco, S. Villa, and B. L. Vu. Don't relax: Early stopping for convex regularization. arXiv preprint https://arxiv.org/abs/1707.05422, 2017.

[46] Y. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady, 269(3):543–547, 1983.

[47] Y. Nesterov. Introductory Lectures on Convex Optimization, volume 87. Springer Science & Business Media, 2004.

[48] A. Neubauer. On nesterov acceleration for landweber iteration of linear ill-posed problems. Journal of Inverse and Ill-posed Problems, 25(3), 2016.

[49] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, 60(1):259–268, 1992.

[50] S. Salzo and S. Villa. Inexact and accelerated proximal point algorithms. Journal of Convex Analysis, 19(4):1167–1192, 2012.

[51] M. Schmidt, N. L. Roux, and F. R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 24, pages 1458–1466. Curran Associates, Inc., 2011.

[52] C. M. Stein. Estimation of the mean of a multivariate normal distribution. The annals of Statistics, pages 1135–1151, 1981.

[53] I. Steinwart and A. Christmann. Support Vector Machines. Springer Science & Business Media, 2008.

[54] W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. Journal of Machine Learning Research, 17(153):1–43, 2016.

[55] S. Villa, S. Salzo, L. Baldassarre, and A. Verri. Accelerated and inexact forward-backward algorithms. SIAM Journal on Optimization, 23(3):1607–1633, 2013.

[56] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67, 2006.

[57] C. Zalinescu. Convex Analysis in General Vector Spaces. World Scientific, 2002.

[58] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2):301–320, 2005.